



FUNDAÇÃO UNIVERSIDADE FEDERAL DE RONDÔNIA – UNIR
DEPARTAMENTO DE MATEMÁTICA E ESTATÍSTICA – DME
CAMPUS DE JI-PARANÁ
CURSO DE ESTATÍSTICA

**MODELOS PARA PREVISÃO EM SÉRIES TEMPORAIS: A
METODOLOGIA BOX E JENKINS NA PRÁTICA**

Ji-Paraná/RO

2017

EVELIM SIEBEN

**MODELOS PARA PREVISÃO EM SÉRIES TEMPORAIS: A
METODOLOGIA BOX E JENKINS NA PRÁTICA**

Trabalho de Conclusão de Curso submetido ao Departamento de Matemática e Estatística da Fundação Universidade Federal de Rondônia, *Campus* de Ji-Paraná, como parte dos requisitos para obtenção do título de Bacharel em Estatística, sob a orientação da Prof^a. Gr^a. Luana Lúcia Alves de Azevêdo.

Ji-Paraná/RO

2017

Dados Internacionais de Catalogação na Publicação
Fundação Universidade Federal de Rondônia
Gerada automaticamente mediante informações fornecidas pelo(a) autor(a)

S571m Sieben, Evelim.

Modelos Para Previsão em Séries Temporais: A Metodologia Box e Jenkins na Prática / Evelim Sieben. -- Ji-Paraná, RO, 2017.

90 f. : il.

Orientador(a): Prof.^a Luana Lúcia Alves de Azevêdo

Trabalho de Conclusão de Curso (Graduação em Estatística) - Fundação Universidade Federal de Rondônia

1. Séries Temporais. 2. Box e Jenkins. 3. Modelos ARIMA. 4. IPCA. I. Azevêdo, Luana Lúcia Alves de. II. Título.

CDU 519.246.8

Bibliotecário(a) Alex Almeida

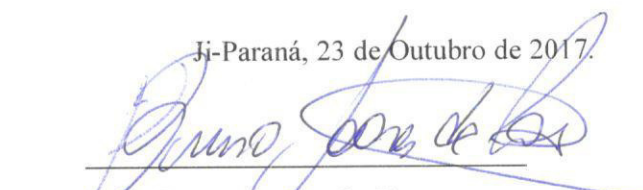
CRB 11.853




ATA DE SESSÃO DE APRESENTAÇÃO E DEFESA DO TRABALHO DE CONCLUSÃO DE CURSO

Aos 23 (vinte e três) dias do mês de outubro de 2017, realizou-se na Sala Laboratório 1 de Estatística, no *Campus* de Ji-Paraná, a Sessão de Apresentação e Defesa do Trabalho de Conclusão de Curso (TCC) da acadêmica **Evelim Sieben**, com o título “**MODELOS PARA PREVISÃO EM SÉRIES TEMPORAIS: A METODOLOGIA BOX & JENKINS NA PRÁTICA**”. Os trabalhos foram iniciados às 16:00 horas pela presidente da Banca Examinadora, aprovada pelo Departamento e constituída por: Prof^o. **Gr. Bruno Soares de Castro** (Universidade Federal de Rondônia - Departamento de Matemática e Estatística), Prof^o. **Dr. Dilson Henrique Ramos Evangelista** (Universidade Federal de Rondônia - Departamento de Matemática e Estatística) e Prof^a. **Gr^a. Luana Lúcia Alves de Azevêdo** (Universidade Federal de Rondônia - Departamento de Matemática e Estatística), sendo esta a orientadora da acadêmica. A Banca Examinadora, tendo decidido aceitar o Trabalho de Conclusão de Curso, após a apresentação passou à arguição pública da acadêmica. Encerrando os trabalhos de arguição às 16:55 horas. A Banca Examinadora deu parecer final **Aprovado**, com a nota 90 (noventa), resultado da média aritmética das notas individuais atribuídas pelos membros da Banca Examinadora. Proclamado o resultado final pela presidente da Banca, foram encerrados os trabalhos. Para constar, lavrou-se a presente Ata, que é assinada pelos membros da Banca Examinadora e pela acadêmica.


Dr. Dilson Henrique Ramos Evangelista
Prof^o. Membro da Banca

Ji-Paraná, 23 de Outubro de 2017.

Gr. Bruno Soares de Castro
Prof. Membro da Banca


Gr^a. Luana Lúcia Alves de Azevêdo
Prof^a. Orientadora e Pres. Banca


Evelim Sieben
Acadêmica

RESUMO

SIEBEN, Evelim. **Modelos para Previsão em Séries Temporais: A Metodologia Box e Jenkins na Prática**. 2017. 90 f. Trabalho de Conclusão de Curso (Bacharelado em Estatística) – Departamento de Matemática e Estatística, Universidade Federal de Rondônia, Ji-Paraná.

Tanto na Economia, quanto na Engenharia e nas Ciências da Natureza, ocorrem fenômenos que dependem da observação de dados em intervalos de tempo durante determinado período. Tratam-se das Séries Temporais. A sua análise tem grande relevância prática, uma vez que, pode revelar através de um modelo adequado, algumas características importantes, as quais auxiliam na previsão de seus valores futuros ou simplesmente descrevem o seu comportamento. A literatura ampara diversos modelos de previsão de séries temporais. Neste trabalho, procura-se reunir os principais aspectos metodológicos de alguns destes modelos, com enfoque nos modelos ARIMA propostos por Box e Jenkins (1970), considerada a metodologia mais importante (SOUZA, L. V., 2006; FIGUEREDO, 2008) e largamente utilizada na análise de séries temporais. Em seguida, a fim de demonstrar uma aplicação prática da Metodologia Box e Jenkins, o estudo ajusta modelos SARIMA à série mensal de dados correspondente ao Índice de Preços ao Consumidor Amplo (IPCA), compreendida no período de janeiro de 2007 a dezembro de 2016. Para tanto a análise esteve pautada no ciclo iterativo proposto por tal metodologia, a saber, nas fases da identificação, estimação, verificação e, finalmente, previsão. Analisando os resultados obtidos, sugeriu-se, em termos de capacidade preditiva, o modelo $SARIMA(1, 0, 0)(0, 1, 1)_{12}$ como modelo mais adequado para os dados deste estudo. Porém, por se tratar de modelos univariados, os modelos SARIMA excluem efeitos de outras variáveis ou eventuais choques que podem ocorrer sobre a inflação mensal. Neste contexto, sugere-se para trabalhos futuros um estudo mais aprofundado sobre a dinâmica do IPCA, a fim de identificar variáveis intrínsecas ao processo, as quais poderão ser incorporadas à modelagem, através do uso de modelos SARIMAX, tornando a previsão mais acurada.

PALAVRAS-CHAVE: Séries Temporais, Box e Jenkins, Modelos ARIMA, IPCA.

ABSTRACT

SIEBEN, Evelim. **Analysis of Time Series: The Box e Jenkins Methodology in Practice.** 2017. 90 f. Monograph (Bachelor of Statistics) - Department of Mathematics and Statistics, Federal University of Rondônia, Ji-Paraná.

In economics, in engineering and natural sciences, phenomena occur that depend of the observation of data at intervals of time during a given period. This collection of observed data is called the Time Series. Its analysis has great practical relevance, since it can reveal through an appropriate model some important characteristics, which help in predicting its future values or simply describe its behavior. The literature supports several models of time series forecasting. In this paper, is described main methodological aspects of some of these models, with focus on ARIMA models proposed by Box e Jenkins (1970), considered the most important methodology (SOUZA, L. V., 2006; FIGUEREDO, 2008) and widely used in the analysis of time series. Then, in order to demonstrate a practical application of the Box e Jenkins Methodology, the study adjusts SARIMA models to the monthly data serie corresponding to the IPCA from January 2007 to December 2016. For this, the analysis was based on the iterative cycle proposed by this methodology, in the identification, estimation, verification and, finally, prediction phase. Analyzing the results, it was suggested, in terms of predictive capacity, the $SARIMA(1,0,0) (0,1,1)_{12}$ model more suitable for the data of this study. But, because they are univariate models, the models SARIMA exclude effects of other variables or possible shocks that may occur on monthly inflation. In this context, a more in-depth study on the dynamics of the IPCA is suggested for future work, in order to identify variables intrinsic to the process, which can be incorporated into the modeling, through the use of SARIMAX models, making the forecast more accurate.

KEYWORDS: Time Series, Box e Jenkins, ARIMA model, IPCA.

LISTA DE QUADROS

Quadro 1.	Resumo das características teóricas da FAC e FACP para os modelos $AR(p)$, $MA(q)$ e $ARMA(p, q)$	48
Quadro 2.	Principais Medidas de Acuracidade.....	56

LISTA DE TABELAS

Tabela 1.	Medidas descritivas da série mensal correspondente ao Índice de Preços ao Consumidor Amplo (IPCA).	60
Tabela 2.	Razão entre a amplitude máxima e a amplitude mínima da série correspondente ao IPCA para série original, série de raiz quadrada e série de logaritmos.	63
Tabela 3.	Coeficientes de variação da série correspondente ao IPCA para série original, série de raiz quadrada e série de logaritmos.	63
Tabela 4.	Resultados dos Testes ADF, PP e KPSS aplicados à série mensal correspondente ao IPCA submetida a uma diferença sazonal.	66
Tabela 5.	Estatísticas dos modelos $SARIMA(p, d, q)(P, D, Q)_{12}$ estimados para a série mensal correspondente ao IPCA.	67
Tabela 6.	Resultado dos p-valores para os teste de Ljung-Box, Box-Pierce, Shapiro-Wilk, Kolmogorov-Smirnov e Jarque-Bera para os modelos ajustados à série mensal correspondente ao IPCA.	68
Tabela 7.	Resultados dos Critérios de Informação AIC, AICC e BIC e das Medidas de Acurácia MAE e RMSE para os modelos ajustados à série mensal correspondente ao IPCA.	68
Tabela 8.	Previsões $h = 8$ para o modelo $SARIMA(1, 0, 0)(0, 1, 1)_{12}$	69

LISTA DE FIGURAS

Figura 1.	Representação gráfica da série original correspondente IPCA compreendida no período de janeiro de 2007 a dezembro de 2016.	61
Figura 2.	Representação gráfica das funções de (1) autocorrelação (FAC) e (2) autocorrelação parcial (FACP) da série original correspondente ao IPCA compreendida no período de janeiro de 2007 a dezembro de 2016.....	62
Figura 3.	Representação gráfica da decomposição da série original correspondente ao IPCA compreendida no período de janeiro de 2007 a dezembro de 2016, nas componentes de tendência, sazonalidade e erro.....	64
Figura 4.	Representação gráfica da série correspondente ao IPCA compreendida no período de janeiro de 2007 a dezembro de 2016, submetida a uma diferenciação sazonal ($D = 1$).....	65
Figura 5.	Representação gráfica das funções de (1) autocorrelação (FAC) e (2) autocorrelação parcial (FACP) da série mensal correspondente ao IPCA, compreendida no período de janeiro de 2007 a dezembro de 2016 submetida à uma diferença sazonal ($D = 1$).	66
Figura 6.	Comparação entre os valores reais da série mensal IPCA e os valores estimados pelos modelos SARIMA no período de janeiro a agosto de 2017.	69
Figura 7.	Ajuste do Modelo SARIMA (1, 0, 0) (0, 1, 1) ₁₂ à série IPCA.....	70
Figura 8.	Representação gráfica da série mensal correspondente ao IPCA compreendida no período de janeiro de 2007 a dezembro de 2016, a série de índices previstos pelo modelo SARIMA (1,0,0) (0,1,1) ₁₂ para $h = 8$ e o intervalo de confiança para as previsões	70
Figura 9.	Representação gráfica da média versus variância da série correspondente ao IPCA compreendida no período de janeiro de 2007 a dezembro de 2016: (1) Série Original, (2) Transformação Logaritmica e (3) Transformação Raiz Quadrada.	77
Figura 10.	Representação Gráfica do IPCA por mês e do IPCA médio por mês.....	78
Figura 11.	Gráfico Sazonal.....	78
Figura 12.	Periodograma da Série referente ao IPCA.	79
Figura 13.	Análise gráfica de resíduos do modelo SARIMA (0,0,1) (1,1,0) ₁₂	80
Figura 14.	Histograma dos resíduos do modelo SARIMA (0,0,1) (1,1,0) ₁₂ com curva normal sobrepota.....	80
Figura 15.	Análise gráfica de resíduos do modelo SARIMA (1,0,0) (1,1,0) ₁₂	81

Figura 16.	Histograma dos resíduos do modelo $SARIMA (1,0,0) (1,1,0)_{12}$ com curva normal sobreposta.	81
Figura 17.	Análise gráfica de resíduos do modelo $SARIMA (1,0,0) (0,1,1)_{12}$	82
Figura 18.	Histograma dos resíduos do modelo $SARIMA (1,0,0) (0,1,1)_{12}$ com curva normal sobreposta.	82

LISTA DE SIGLAS E ABREVIATURAS

ADF	Dickey Fuller Aumentado
AIC	Critério de Informação de Akaike
AR	Autorregressivo
ARCH	Autorregressivo Condicional Heterocedástico
ARFIMA	Autorregressivo Fracionário Integrado de Médias Móveis
ARIMA	Autorregressivo Integrado de Médias Móveis
ARIMAX	Autorregressivo Integrado de Médias Móveis com Entrada de Variáveis Exógenas
ARMA	Autorregressivo de Médias Móveis
ARMAX	Autorregressivo de Médias Móveis com Entrada de Variáveis Exógenas
BIC	Critério de Informação Bayesiano
CH	Canova-Hansen
DF	Dickey Fuller
EGARCH	Exponencial Autorregressivo com Heterocedasticidade Condicional
FAC	Função de Autocorrelação
FACP	Função de Autocorrelação Parcial
FACV	Função de Autocovariância
GARCH	Generalizado Autorregressivo Condicional Heterocedástico
IBGE	Instituto Brasileiro de Geografia e Estatística
IDE	Integrated Developement Environment
IPCA	Índice de Preços ao Consumidor Amplo
IPEA	Instituto de Pesquisa Econômica Aplicada
KPSS	Kwiatkowski-Phillips-Schmidt-Shin
MA	Médias Móveis
MAE	Erro Absoluto Médio
MAPE	Erro Percentual Médio Absoluto

MLD	Modelos Lineares Dinâmicos
MMV	Método de Máxima Verossimilhança
MQO	Mínimos Quadrados Ordinários
MSE	Erro Quadrático Médio
MVE	Modelos de Volatilidade Estocástica
OCSB	Osborn-Chui-Smith-Birchenhall
PP	Phillips-Perron
QQ-Plot	Gráfico Quantil Quantil
SARFIMA	Autorregressivo Fracionário Integrado de Médias Móveis Sazonal
SARIMA	Autorregressivo Integrado de Médias Móveis Sazonal
SARIMAX	Autorregressivo Integrado de Médias Móveis Sazonal com Entrada de Variáveis Exógenas
STAR	Autorregressivo com Transição Suave
TAR	Autorregressivo Limiar
TARCH	Autorregressivo com Heterocedasticidade Condicional e Limiar

SUMÁRIO

1 INTRODUÇÃO	14
2 REFERENCIAL TEÓRICO	15
2.1 SÉRIES TEMPORAIS	15
2.1.1 Breve Histórico da Modelagem de Séries Temporais	15
2.1.2 Definição e Suposições	16
2.1.3 Algumas Definições Matemáticas	18
2.1.4 Decomposição de Uma Série Temporal.....	20
2.2 MODELOS DE PREVISÃO	22
2.2.1 Modelos de Suavização Exponencial	23
2.2.1.1 Médias Móveis Simples	23
2.2.1.2 Médias Móveis Ponderadas	24
2.2.1.3 Suavização Exponencial Simples	24
2.2.1.4 Suavização Exponencial de Holt	26
2.2.1.5 Suavização Exponencial Sazonal de Holt-Winters	27
2.2.2 Modelos ARIMA	28
2.2.2.1 Modelos Autorregressivos (AR).....	29
2.2.2.2 Modelos de Médias Móveis (MA)	30
2.2.2.3 Modelos Autorregressivos e de Médias Móveis (ARMA)	30
2.2.2.4 Modelos Autorregressivos Integrados de Médias Móveis (ARIMA).....	31
2.2.2.5 Modelos Autorregressivos Integrados de Médias Móveis Sazonal (SARIMA)....	32
.....	32
2.3 PROCEDIMENTOS ÚTEIS NA MODELAGEM DE SÉRIES TEMPORAIS	33
2.3.1 Testes de Raiz Unitária Simples.....	33
2.3.1.1 Teste de Dickey-Fuller	33
2.3.1.2 Teste de Phillips-Perron	35
2.3.1.3 Teste de Kwiatkowski, Philips, Schmidt e Shin	35
2.3.2 Testes para Tendência	36
2.3.2.1 Teste de Wald-Wolfowitz.....	36
2.3.2.2 Teste de Cox-Stuart	37
2.3.2.3 Teste Baseado no Coeficiente de Correlação de Spearman	37
2.3.2.4 Teste de Mann-Kendall	38
2.3.3 Testes de Raiz Unitária Sazonal.....	39
2.3.3.1 Teste HEGY	39
2.3.3.2 Teste de Osborn-Chui-Smith-Birchenhall	40
2.3.3.3 Teste de Canova-Hansen	40

2.3.4 Testes para Sazonalidade.....	42
2.3.4.1 Teste de Kruskal-Wallis	42
2.3.4.2 Teste de Friedman	43
2.3.4.3 Teste de Fisher.....	43
2.3.5 Transformações	44
2.3.5.1 Diferenciação.....	45
2.3.5.2 Transformação Box-Cox	45
2.4 METODOLOGIA BOX e JENKINS	46
2.4.1 Identificação.....	47
2.4.1.1 Função de Autocorrelação e Função de Autocorrelação Parcial.....	47
2.4.2 Estimação	49
2.4.3 Verificação	49
2.4.3.1 Análise de Resíduos	50
2.4.3.2 Testes de Adequação do Modelo.....	50
2.4.3.2.1 Teste de Autocorrelação Residual	50
2.4.3.2.2 Teste de Box-Pierce.....	51
2.4.3.2.3 Teste de Ljung-Box.....	51
2.4.3.3 Testes de Normalidade dos Resíduos	52
2.4.3.3.1 Teste de Shapiro-Wilk	52
2.4.3.3.2 Teste de Kolmogorov-Smirnov	52
2.4.3.3.3 Teste de Jarque-Bera	53
2.4.3.4 Critérios de Seleção de Modelos	54
2.4.3.4.1 Critério de Akaike	54
2.4.3.4.2 Critério de Akaike Corrigido.....	54
2.4.3.4.3 Critério de Informação Bayesiano	55
2.4.3.5 Medidas de Acuracidade	55
2.4.4 Previsão	56
2.4.4.1 Atualização das Previsões	57
3 METODOLOGIA.....	57
3.1 CONSIDERAÇÕES SOBRE O IPCA	59
4 ANÁLISE E DISCUSSÃO DOS RESULTADOS	60
5 CONCLUSÃO.....	71
REFERÊNCIAS	73
APÊNDICE A	77
APÊNDICE B.....	78
APÊNDICE C	80
APÊNDICE D	83

1 INTRODUÇÃO

A previsão é imprescindível no processo de planejamento, na tomada de decisão e no controle, isto em qualquer domínio de atividade. A incerteza associada à maioria dos fenômenos temporais não permite o conhecimento exato do seu comportamento futuro, tendo-se que recorrer às previsões. Uma previsão consiste numa tentativa de explicar o comportamento futuro a partir das condições atuais, utilizando-se dados históricos do passado, tendo por finalidade facilitar planejamento e a tomada de decisões por meio de informações probabilísticas.

Segundo Souza, F. M., (2006), a previsão não constitui um fim em si, mas um meio de fornecer informações e subsídios para uma consequente tomada de decisões, visando atingir determinados objetivos. No processo de análise, pressupõe-se que as condições no presente, mesmo as que se ignoram, não vão sofrer alterações no futuro.

Segundo Cortez (2002):

O desejo de compreender o passado e prever o futuro impulsiona a procura por leis que expliquem o comportamento de certos fenômenos ou acontecimentos. Se as equações que os determinam são conhecidas, é possível utilizar as mesmas para prever o resultado de um determinado experimento, desde que sejam conhecidas as condições iniciais. Na ausência de regras que definam o comportamento de um sistema, procura-se determinar o seu comportamento futuro a partir de observações concretizadas no passado. (CORTEZ, 2002, pág. 2).

Neste contexto, as séries temporais constituem fonte de informação para o conhecimento da realidade e para pesquisa científica em diversos setores da sociedade. Sua análise permite a construção de modelos de previsão que necessitam basicamente de uma série de valores observados ao longo do tempo. A título de exemplo de séries temporais, convém citar séries financeiras, econométricas, meteorológicas, oceanográficas, pluviométricas, séries de demanda de energia, custo de vida, índice de produção, etc.

A análise de séries temporais tem por finalidade identificar e descrever os padrões de comportamento a partir da relação causal com o tempo, o qual exerce influência sobre os dados, auxiliando na compreensão do passado e na predição do futuro, permitindo a tomada de decisões devidamente fundamentadas.

Nas previsões modeladas através de séries temporais, a predição tem por base o estudo dos dados históricos da variável de interesse, isto é, baseiam-se no estudo do comportamento de determinada variável num dado período de tempo. Segundo Relvas (1998), isto ocorre porque um modelo de séries temporais, sempre assume que algum padrão ou combinação de padrões é recorrente através do tempo. De acordo com a autora, esse tipo de análise não

indicará todas as respostas válidas para explicar o futuro, mas certamente será extremamente útil para o processo de previsão.

Os métodos estatísticos para análise de séries temporais estão amplamente desenvolvidos na literatura. Encontram-se disponíveis diversos modelos probabilísticos, entretanto, cabe ao usuário decidir o procedimento mais adequado à sua análise, uma vez que cada modelo exige suposições às características dos dados. Segundo Morettin e Toloi (2004),

Qualquer que seja a classificação que façamos para os modelos de séries temporais, podemos considerar um número muito grande de modelos diferentes para descrever o comportamento de uma série particular. A construção destes modelos depende de vários fatores, tais como o comportamento do fenômeno ou o conhecimento *a priori* que temos de sua natureza e do objetivo da análise. (MORETTIN e TOLOI, 2004, pág. 21).

Embora seja vasta a lista de técnicas para realizar a previsão de séries temporais, todas compartilham de um erro de previsão. Dessa maneira o analista é desafiado a minimizar este erro, construindo um modelo que se ajuste adequadamente aos dados utilizando um número parcimonioso de parâmetros. A construção desse modelo não é tarefa simples e pode conduzir a uma previsão pouco eficiente.

De acordo com Cortez (2002) a utilização de qualquer metodologia deve ter por objetivo minimizar o erro de previsão através da construção de um modelo que melhor se ajuste aos dados, fazendo-se uso de um número parcimonioso de parâmetros.

Dada a importância do estudo de séries temporais, este trabalho tem por objetivo apresentar e discorrer sobre alguns modelos utilizados em sua análise, dando enfoque à metodologia Box e Jenkins e, a título de ilustração, aplicar a referida metodologia à série de dados referentes ao Índice de Preços ao Consumidor Amplo (IPCA), compreendida no período de janeiro de 2007 a agosto de 2017.

2 REFERENCIAL TEÓRICO

2.1 SÉRIES TEMPORAIS

2.1.1 Breve Histórico da Modelagem de Séries Temporais

A modelagem de séries temporais surgiu da necessidade de se acompanhar e analisar dados coletados em sucessivos intervalos de tempo, isto porque um conjunto de dados, por si só, não permite a previsão de valores futuros. Para isso, é necessário que sejam utilizados

métodos estatísticos de previsão, os quais envolvem cálculos e procedimentos específicos aplicáveis a esta série de dados.

Em geral, os métodos estatísticos de previsão de séries temporais buscam identificar um padrão de comportamento da série e utilizá-lo para prever os valores futuros. Segundo Souza, L. V., (2006), até 1920, a previsão era efetuada através da extrapolação simples de um valor global ajustado em função do tempo. As previsões eram calculadas apenas ao redor da média. A partir de 1927, Yule desenvolveu o modelo Autorregressivo (AR), no qual o valor previsto dependia de valores passados. Nos próximos vinte anos seguintes os pesquisadores acreditaram que o comportamento de uma série dependia apenas dos modelos lineares e do ruído, no entanto, pesquisas posteriores mostraram que os modelos lineares não supriam completamente a análise de séries temporais, uma vez que, a maior parte das séries encontradas na prática apresenta alguma forma de não estacionariedade (MORETTIN e TOLOI, 2004).

Os estudos prosseguiram e na década de 1950 passou-se a utilizar o Alisamento Exponencial, de fácil utilização e que exige poucos cálculos. Finalmente, na década de 1970 surgiu a metodologia Box e Jenkins, a qual é capaz de fornecer modelos ARIMA que descrevem, com certa precisão e de forma parcimoniosa, o mecanismo gerador da série temporal. Trata-se de uma técnica completa de identificação do modelo que melhor se ajusta aos dados, de estimação dos parâmetros e da verificação da validade do modelo empregado (SOUZA, L. V., 2006). Esta técnica propõe um modelo probabilístico capaz de fornecer previsões, porém apresenta algumas restrições, tais como a amostra deve conter pelo menos 50 observações.

Ao longo do tempo, outras técnicas foram testadas. Em 1980, Thiele contribuiu relacionando a formulação e análise de um método para uma série temporal constituído de uma soma de componentes de regressão, movimento Browniano, que recebeu o nome de Filtro de Kalman. A partir disso, foram desenvolvidas inúmeras técnicas como tratamento de dados faltantes, para os métodos de suavização exponencial e ajuste sazonal, entre outros; todas com o intuito de se efetuar previsões mais precisas. Novos modelos também foram propostos. Citam-se entre eles as Redes Neurais Artificiais, os Algoritmos Genéticos e os Algoritmos Evolutivos, os quais têm apresentado bons resultados.

2.1.2 Definição e Suposições

Existem diversas definições para série temporal.

Segundo Morettin e Toloi (2004), uma série temporal é qualquer conjunto de observações ordenadas no tempo.

Carvalho (2006) traz a seguinte definição: “uma série temporal, também chamada série cronológica ou histórica, pode ser definida como a realização de um processo estocástico, sendo este uma sequência de variáveis aleatórias no tempo”. Segundo Morettin e Toloi (2004), um processo estocástico pode ser definido como um processo controlado por leis probabilísticas.

Gujarati (2006) define série temporal como um conjunto de observações dos valores que uma variável assume em diferentes momentos de tempo.

Miranda (2014) diz que uma série temporal é um conjunto de observações de uma determinada variável feitas em períodos sucessivos de tempo, ao longo de um determinado intervalo.

Qualquer que seja a definição, a hipótese fundamental é a mesma: o tempo é o principal condutor do fenômeno em estudo.

O objetivo da análise de uma série temporal é identificar padrões não aleatórios nesta, de maneira que a observação deste comportamento passado permita fazer previsões sobre o futuro. Segundo Ehlers (2005), a característica mais importante desse tipo de dados é que as observações vizinhas são dependentes. O objetivo, portanto, no seu estudo é analisar e modelar esta dependência.

Outro fator que se deve levar em conta é a ordem das observações. Variar essa ordem pode modificar a informação contida na série (MARTIN, 2015). Os modelos de séries temporais se baseiam na ideia de que à proporção que as observações se afastam no tempo do momento presente, estas perdem importância no processo de construção das projeções (GARBER, 1995), isto é, o que acontece no futuro próximo dependerá, em grande parte, do que aconteceu ou está acontecendo agora. Em outras palavras, a “memória temporal” existente entre as observações próximas é um aspecto essencial quando se deseja aplicar técnicas de previsão.

Uma série temporal pode ainda ser classificada como contínua, quando as observações são feitas continuamente no tempo; ou discreta quando as observações são feitas em tempos específicos, geralmente equiespaçados (EHLERS, 2005). No caso de um conjunto de dados com valores contínuos, este deve ser discretizado no tempo e com intervalos iguais (MORETTIN e TOLOI, 2004).

Uma das suposições mais frequentes que se faz a respeito de uma série temporal é a de que ela é estacionária, ou seja, se desenvolve aleatoriamente ao redor de uma média constante

(MORETTIN e TOLOI, 2004). Desta forma, a maioria dos procedimentos de análise estatística de séries temporais supõe que estas sejam estacionárias. Porém, na prática, a maior parte das séries reais apresenta alguma forma de não estacionariedade. Logo, durante o processo, será necessário transformar os dados originais, caso estes não formem uma série estacionária.

De acordo com o modelo clássico, uma série temporal é composta de quatro componentes: tendência, ciclo, sazonalidade e ruído (variações aleatórias). O processo de construção de valores previstos para a série é realizado através da reunificação de cada um desses componentes (HENNING *et al*, 2015). Assim, conhecendo seus componentes e pressupondo-se a existência de algum padrão geral e não aleatório de comportamento esses componentes poderão ser descritos por funções matemáticas e, por meio delas, será possível prever o comportamento futuro da variável estudada (RELVAS, 1998).

2.1.3 Algumas Definições Matemáticas

Em termos matemáticos uma série temporal pode ser definida como qualquer conjunto Z de observações ordenadas no tempo:

$$Z = Z_t, \quad t = 1, 2, \dots, N, \quad (2.1)$$

em que t é o índice de tempo e N é o número de observações.

O que se chama de série temporal é parte de uma trajetória, dentre muitas que poderiam ter sido observadas dentro de um processo estocástico. Processo estocástico nada mais é do que o conjunto de todas essas possíveis trajetórias (séries) que poderiam ser observadas (MORETTIN e TOLOI, 2004).

Formalmente, um processo estocástico pode ser definido como: Seja T um conjunto arbitrário. Um processo estocástico é uma família $Z = \{Z(t), t \in T\}$, tal que, para cada $t \in T$, $Z(t)$ é uma variável aleatória. Nestas condições, um processo estocástico é uma família de variáveis aleatórias, supostamente definidas um mesmo espaço de probabilidades (MORETTIN e TOLOI, 2004).

A função *media* de Z é dada por:

$$\mu(t) = E\{Z(t)\} = \int_{-\infty}^{\infty} zf(z; t)dz, \quad (2.2)$$

podendo ser estimada pela equação:

$$\bar{Z}_t = \frac{\sum_{t=1}^n Z_t}{n}. \quad (2.3)$$

A função de *autocovariância* (FACV) do processo estocástico, $Z(w, t)$, é o momento central conjunto das variáveis aleatórias Z_t e Z_{t+k} , calculada por:

$$\hat{\gamma}_k = E[(Z_t - \mu(t))(Z_{t+k} - \mu(t+k))], \quad (2.5)$$

em que z_t é a observação no instante t . Se o processo for estacionário, então:

$$\hat{\gamma}_k = E[(Z_t - \mu)(Z_{t+k} - \mu)]. \quad (2.6)$$

O estimador para a *autocovariância* é dado pela seguinte equação:

$$\bar{\gamma}_k = \frac{1}{n} \sum_{t=1}^n (z_t - \bar{z})(z_{t+k} - \bar{z}). \quad (2.7)$$

A função *variância* de um processo estocástico estacionário, representada por $Var[Z(t)] = E(Z_t - \mu)^2 = \sigma^2 = \gamma_0$, é obtida da função de *autocovariância* (FACV) quando $k = 0$ e pode ser calculada por meio do seu estimador:

$$\hat{\gamma}_0 = \frac{1}{n} \sum_{t=1}^n (z_t - \bar{z})^2 = \hat{\sigma}_z^2. \quad (2.4)$$

Portanto, considerando um processo, $Z(w, t)$ e os instantes t_1, t_2, \dots, t_{n_1} pertencentes ao conjunto de instantes T , este é denominado estacionário no sentido amplo, se para qualquer número n de variáveis aleatórias Z_1, Z_2, \dots, Z_n a função da média, $\mu(t) = E[Z_t] = \mu$ e $E[Z_t^2] < \infty$, ou seja, a esperança é finita e a função de *autocovariância*, $\gamma(t_1, t_k) = E[(Z_{t+k} - \mu)(Z_t - \mu)] = \gamma_k$ é uma função da defasagem k .

Ehlers (2005) afirma que a função de *autocorrelação* teórica (FAC) de um processo estocástico estacionário é uma ferramenta importante para acessar suas propriedades. A função de *autocorrelação* de um processo $Z(w, t)$, representada por ρ_k , para a ordem k , é a razão, $\rho_k = \frac{\gamma_k}{\gamma_0}$ entre a *autocovariância* de ordem k , γ_k , e a variância do processo, γ_0 . O estimador da função é dado por:

$$\hat{\rho}_k = \frac{\sum_{t=1}^n (z_t - \bar{z})(z_{t+k} - \bar{z})}{\sum_{t=1}^n (z_t - \bar{z})^2}. \quad (2.8)$$

A função de autocorrelação parcial (FACP), por sua vez, pode ser definida como a sequência de correlações entre $(Z_t \text{ e } Z_{t-1}), (Z_t \text{ e } Z_{t-2}), (Z_t \text{ e } Z_{t-3})$ e assim por diante, desde que os efeitos de defasagens anteriores sobre Z_t permaneçam constantes. A FACP é calculada como o valor do coeficiente ϕ_{kk} na equação: $Z_t = \phi_{k1}Z_{t-1} + \phi_{k2}Z_{t-2} + \phi_{k3}Z_{t-3} + \dots + \phi_{kk}Z_{t-k} + e_t$ (CRUZ, 2008).

Como já dito anteriormente, uma classe de processos estocásticos importante na análise de séries temporais é a classe dos processos estacionários. Segundo Morettin e Toloi (2004), um processo Z é estacionário se ele se desenvolve no tempo de modo que a escolha de uma origem dos tempos não é importante. Isto é, as características de $Z_t(t + \tau)$, para todo τ , são as mesmas de $Z(t)$.

Um processo estocástico é dito estritamente estacionário se todas as distribuições finito-dimensionais $F(z_1, \dots, z_n; t_1, \dots, t_n) = P\{Z(t_1) \leq z_1, \dots, Z(t_n) \leq z_n\}$, permanecem as mesmas sob translações do tempo, ou seja, $F(z_1, \dots, z_n; t_1 + \tau, \dots, t_n + \tau) = F(z_1, \dots, z_n; t_1, \dots, t_n)$, para quaisquer t_1, \dots, t_n, τ de T . Logo, a média e a variância são constantes. Porém, na prática, é muito difícil usar a definição de estacionariedade estrita e costuma-se definir estacionariedade de uma forma menos restrita (EHLERS, 2005; MORETTIN e TOLOI, 2004).

Se o modelo for estacionário na média e na variância e a correlação depender apenas do número de intervalo de tempo que as separam (defasagem), o modelo é dito ser fracamente estacionário. Diz-se que um processo estocástico é fracamente estacionário ou estacionário de segunda ordem se, e somente se:

- i) $\{Z(t)\} = \mu(t) = \mu$, constante para todo $t \in T$;
- ii) $E\{Z^2(t)\} < \infty$, para todo $t \in T$;
- iii) $\gamma(t_1, t_2) = Cov\{Z(t_1), Z(t_2)\}$ é uma função de $|t_1 - t_2|$.

2.1.4 Decomposição de Uma Série Temporal

Conforme já visto, a maneira usual de analisar uma série temporal é através da sua decomposição nas componentes de tendência, ciclo e sazonalidade, o que possibilita analisar cada uma delas separadamente. O objetivo não se resume à simples decomposição. Após a remoção sistemática de cada componente, a começar pela tendência, eles são analisados isoladamente e, depois, a série é recomposta para que se possa apurar os desvios entre a série

original e a série calculada, a fim de entender as variações passadas, prever os movimentos futuros e proceder aos testes e controle do processo para fins de melhoria (RELVAS, 1998).

Tendência é um comportamento de longo prazo no nível médio da série (EHLERS, 2005). Ciclos são flutuações nos valores da variável com duração superior a um ano e que se repetem com certa periodicidade. Sazonalidade refere-se a flutuações nos valores da variável com duração inferior a um ano e que se repetem regularmente no tempo. Uma série pode apresentar ainda variações irregulares, denominadas ruído ou erro, que são flutuações inexplicáveis causadas por eventos imprevisíveis e não periódicos.

Considerando, portanto, que uma série temporal é resultante da combinação desses fatores, o processo desses componentes pode ser aditivo, conforme abaixo:

$$Z_t = T_t + S_t + C_t + R_t \quad (2.9)$$

ou multiplicativo:

$$Z_t = T_t \times S_t \times C_t \times R_t, \quad (2.10)$$

em que T_t é uma componente de tendência, S_t é uma componente sazonal, C_t é uma componente cíclica e R_t é uma componente aleatória ou ruído.

Segundo Relvas (1998),

O modelo aditivo deve ser adotado quando os componentes da série atuam isoladamente, ou seja, de modo absoluto e independente entre si. Já o modelo multiplicativo, que considera uma série temporal como resultante do produto de seus componentes, deve ser adotado quando os componentes da série atuem de modo proporcional às suas respectivas forças (RELVAS, 1998).

Ainda de acordo com a autora,

A escolha entre um e outro deve estar fundamentada, basicamente, na sensibilidade das variações sazonais em relação ao próprio fenômeno e, sendo apurada uma regularidade aritmética, deve-se adotar o modelo aditivo, caso contrário, o multiplicativo (RELVAS, 1998).

Segundo Morettin e Toloi (2004), se os intervalos sazonais variam com a tendência, um modelo mais adequado é o multiplicativo. Relvas (1998) cita Milone e Angelinis (1995), os quais afirmam que, em termos práticos, o modelo multiplicativo é o que mais se adequa em 75% dos casos.

No entanto, o melhor modelo dependerá dos dados da própria série, das características inerentes ao problema.

2.2 MODELOS DE PREVISÃO

Segundo Morettin e Toloi (2004), a análise de séries temporais pode ter os seguintes objetivos:

- a) Investigar o mecanismo gerador da série temporal; b) fazer previsões de valores futuros da série; c) descrever apenas o comportamento da série; d) procurar periodicidades relevantes nos dados. (MORETTIN e TOLOI, 2004, pág. 3).

De acordo com o autor, em todos os casos, os modelos construídos devem ser simples e parcimoniosos (no sentido que o número de parâmetros envolvidos deve ser o menor possível) e, se possível, sua utilização não deve apresentar dificuldades às pessoas interessadas em manipulá-los.

Para Moreira (2013), os objetivos podem ser os seguintes:

- a) Descrição: descrever propriedades da série, tais como o padrão de tendência, a existência de variações sazonais ou cíclicas, a existência de *outliers* ou alterações estruturais decorrentes das mudanças de tendências ou de sazonalidades; b) Explicação: usar a variação de uma série para explicar a variação de outra série, quando são obtidas observações de duas ou mais variáveis; c) Predição: prever valores futuros com base em valores anteriormente registrados; d) Controle: analisar as origens de falhas para melhorar processos e assim obter um maior rigor, quando os valores da série temporal medem a qualidade de uma atividade. (MOREIRA, 2013, pág. 8).

Ao longo do tempo, diferentes modelos estatísticos foram desenvolvidos para auxiliar na tarefa de previsão de séries temporais. Entre eles têm-se modelos lineares como os Modelos de Suavização Exponencial (Médias Móveis Simples, Médias Móveis Ponderadas, Suavização Exponencial Simples, Suavização Exponencial de Holt, Suavização Exponencial de Holt-Winters) e os Modelos Autorregressivos Integrados de Médias Móveis (AR, MA, ARMA, ARIMA, SARIMA); modelos com entrada de variáveis exógenas (ARMAX, ARIMAX, SARIMAX); modelos lineares dinâmicos (MLD); modelos estruturais; modelos não-lineares (ARCH, GARCH, EGARCH, TARCH, modelos de volatilidade estocástica (MVE), modelos bilineares e modelos Autorregressivos com Limiar (TAR, STAR) (MORETTIN e TOLOI, 2004); modelos com memória longa (ARFIMA, SARFIMA) e modelos multivariados. Há ainda tecnologias de inteligência computacional como Redes Neurais, Lógica Nebulosa e Algoritmos Genéticos, as quais promovem a criação de tecnologias avançadas de previsão.

A seguir serão apresentados alguns dos modelos lineares citados acima.

2.2.1 Modelos de Suavização Exponencial

Os modelos de suavização exponencial são assim conhecidos por assumirem que os valores extremos da série representam a aleatoriedade e, assim, por meio da suavização desses extremos, pode-se identificar o padrão básico desta (MORETTIN e TOLOI, 2004). Basicamente, consistem na aplicação de um conjunto de pesos desiguais aos valores passados da série temporal, sendo que tais pesos decaem de forma exponencial da mais recente para a mais distante observação.

São utilizados em larga escala em virtude da sua simplicidade, do seu baixo custo e precisão razoável, porém tem como desvantagem a dificuldade em determinar os valores mais apropriados das constantes de suavização.

2.2.1.1 Médias Móveis Simples

Este método é indicado para previsões de curto prazo em séries para as quais as componentes de tendência e sazonalidade são inexistentes ou possam ser desprezadas. O nome média móvel é utilizado porque, a cada período, a observação mais antiga é substituída pela mais recente, calculando-se uma média nova (MORETTIN e TOLOI, 2004).

Consiste em calcular a média aritmética das r observações mais recentes para prever o valor do próximo dado.

Desta forma:

$$M_t = \frac{Z_t + Z_{t-1} + \dots + Z_{t-r+1}}{r}. \quad (2.11)$$

Segundo Chambers, Mullick, Smith (1971); Archer (1980); Makridakis, Wheelwright, Hyndman (1998), o número de observações r em cada cálculo da média permanece constante e é estipulado de maneira a tentar eliminar da melhor forma possível as componentes de tendência e sazonalidade (*apud* CORDEIRO *et al*, 2015).

Segundo Morettin e Toloi (2004), o método tem como principais vantagens:

- a) Simples aplicação;
- b) É aplicável quando se tem um número pequeno de observações;
- c) Permite uma flexibilidade grande devido à variação de r de acordo com o padrão da série.

Ainda de acordo com os autores, suas desvantagens são:

- a) Deve ser utilizado somente para séries estacionárias,
- b) Necessidade de armazenar pelo menos $(r - 1)$ observações;
- c) Dificuldade em determinar o valor de r .

Os autores ressaltam ainda que tal método não é utilizado com frequência, pois há outros métodos de suavização exponencial que possuem as citadas vantagens e mais algumas, o que os tornam mais atraentes.

2.2.1.2 Médias Móveis Ponderadas

Também adequado para séries que não apresentam tendência ou sazonalidade, neste método, cada observação na média é ponderada igualmente. Tipicamente, as observações mais recentes carregam pesos maiores na média. Desta forma:

$$M_t = w_t Z_t + w_{t-1} Z_{t-1} + \dots + w_{t-r+1} Z_{t-r+1}, \quad (2.12)$$

em que:

w_i = o peso para a observação i ;

$\sum w_i = 1$, a soma dos pesos é 1.

Um pouco mais complicado de se construir, o modelo média móvel ponderada obterá um resultado melhor.

Umas das desvantagens de se usar esse método para previsão é que durante o cálculo é atribuído o mesmo peso à média das observações, ao passo que o esperado seria que as observações mais recentes tivessem maior peso por serem consideradas uma melhor indicação do futuro.

2.2.1.3 Suavização Exponencial Simples

A suavização exponencial simples é largamente utilizada atualmente, por sua simples aplicação e capacidade de ajustes em relação à acurácia. Trata-se de um filtro estatístico, no qual as entradas são os dados simples a partir de um processo estocástico e a saída são estimativas de uma média que varia com o tempo. Enquanto na média móvel simples, as observações passadas são ponderadas igualmente, esse método atribui pesos decrescendo exponencialmente à medida que as observações vão se tornando mais distantes. Ou seja, às

observações mais recentes são dadas relativamente mais peso na previsão do que as observações mais antigas.

Este método pode ser utilizado para séries estacionárias e sua representação e pode ser descrita matematicamente por (MORETTIN e TOLOI, 2004):

$$\bar{Z}_t = \alpha Z_t + (1 - \alpha)\bar{Z}_{t-1}, \bar{Z}_0 = Z_1, t = 1, \dots, N; \quad (2.13)$$

ou

$$\bar{Z}_t = \alpha \sum_{k=0}^{t-1} (1 - \alpha)^k Z_{t-k} + (1 - \alpha)^t \bar{Z}_0, t = 1, \dots, N, \quad (2.14)$$

em que \bar{Z}_t é denominado valor exponencialmente suavizado e α é a constante de suavização, $0 \leq \alpha \leq 1$.

O valor dessa constante α é arbitrário. Pode-se determinar o melhor valor para esta através de métodos iterativos para minimizar alguma medida da qualidade da previsão. Dessa maneira seleciona-se inicialmente um valor aleatório para a constante, a partir do qual previsões são geradas. Comparam-se os valores previstos com os reais, e calcula-se a média do quadrado das diferenças entre os mesmos. O parâmetro que minimiza essa média é utilizado no modelo final. Segundo Morettin e Toloi (2004), quanto menor for o valor de α mais estáveis serão as previsões finais, uma vez que a utilização de baixo valor de α implica que pesos maiores serão dados às observações passadas e, conseqüentemente, qualquer flutuação aleatória, no presente, exercerá um peso menor no cálculo da previsão.

A previsão de todos os valores futuros é dada pelo último valor exponencialmente suavizado, isto é:

$$\hat{Z}_t(h) = \bar{Z}_t, \forall h > 0, \quad (2.15)$$

$$\hat{Z}_t(h) = \alpha Z_t + (1 - \alpha)\hat{Z}_{t-1}(h + 1), \quad (2.16)$$

que pode ser interpretada como uma equação de atualização da previsão, quando se tiver uma nova observação (MORETTIN e TOLOI, 2004). A previsão reduz o problema de armazenagem de observações, pois pode ser calculada utilizando apenas a observação mais recente, a previsão imediatamente anterior e o valor de α .

Ainda de acordo com os autores, esse método como qualquer outro oferece vantagens e desvantagens. Suas vantagens são:

- a) Fácil entendimento;
- b) Aplicação não dispendiosa;
- c) Grande flexibilidade permitida pela variação da constante de suavização α ;
- d) Necessidade de armazenar somente Z_t , \bar{Z}_t e α ; e
- e) O valor de $\alpha = 2/(r - 1)$ fornece previsões semelhantes ao método Médias Móveis Simples com parâmetro r (MONTGOMERY e JOHNSON, 1976 *apud* MORETTIN e TOLOI, 2004).

Sua principal desvantagem é a dificuldade em determinar o valor mais apropriado da constante de suavização.

2.2.1.4 Suavização Exponencial de Holt

O método de suavização exponencial simples não é indicado para séries que apresentam tendência, uma vez que fornece previsões que subestimam ou superestimam os valores reais. Dessa forma, a fim de evitar esse erro, pode-se aplicar a Suavização Exponencial de Holt. Este método é semelhante, a princípio, a suavização simples. A diferença é que, em vez de suavizar só o nível, utiliza uma nova constante de suavização para modelar a tendência da série (MORETTIN e TOLOI, 2004). De acordo com os autores, os valores do nível e da tendência da série, no instante t serão, respectivamente, estimados por:

$$\bar{Z}_t = AZ_t + (1 - A)(\bar{Z}_{t-1} + \hat{T}_{t-1}), 0 < A < 1 \text{ e } t = 2, \dots, N, \quad (2.17)$$

$$\hat{T}_t = C(\bar{Z}_t - \bar{Z}_{t-1}) + (1 - C)\hat{T}_{t-1}, 0 < C < 1 \text{ e } t = 2, \dots, N. \quad (2.18)$$

A e C serão denominadas constantes de suavização. As fórmulas apresentadas modificam estimativas prévias quando uma nova observação é obtida, assim como em todos os métodos de suavização.

A previsão para o valor de Z_{t+h} , com origem em t é dada por:

$$\hat{Z}_t(h) = \bar{Z}_t + h\hat{T}_t, \forall h > 0, \quad (2.19)$$

isto é, a previsão é feita adicionando-se ao valor básico (\bar{Z}_t) a tendência multiplicada pelo número de passos à frente que se deseja prever (h).

A determinação das constantes de suavização é semelhante ao procedimento adotado para determinar a constante do método de suavização exponencial simples. As vantagens são

semelhantes às do citado método. A desvantagem principal, analogamente, é a dificuldade em determinar os valores mais apropriados das duas constantes de suavização.

2.2.1.5 Suavização Exponencial Sazonal de Holt-Winters

O método de suavização exponencial de Holt-Winters considera na série até três coeficientes de alisamento: nível, tendência e sazonalidade, além de um componente residual aleatório (TÍBULO, 2014). É indicado para previsões de curto prazo e assim, como os demais métodos de alisamento, é de simples aplicação, baixo custo, e possui a capacidade de ajustamento automático rápido à mudanças na série, além de possuir boa precisão.

De acordo com as características da série considerada, há dois tipos de procedimento a serem adotados, os quais podem resultar em dois modelos diferentes: o multiplicativo e o aditivo. Tíbulo (2014), citando Nascimento (2013), afirma que quando a amplitude da série estudada varia de acordo com a sua evolução ou aumenta à medida que o nível da série aumenta, o modelo mais adequado é o multiplicativo. Já nos casos em que o comportamento da série é homogêneo, os melhores resultados serão dados pelo modelo aditivo.

Ambos os procedimentos, porém, irão requerer a determinação de três constantes de suavização, cada uma delas associadas a uma das componentes da série: tendência, nível e sazonalidade.

Segundo Morettin e Tolo (2004), o procedimento mais utilizado considera a sazonalidade como sendo multiplicativa e a tendência permanece aditiva. Desta forma:

$$Z_t = \mu_t F_t + T_t + \varepsilon_t, t = 1, \dots, N. \quad (2.20)$$

As três equações de suavização são dadas abaixo:

$$\hat{F}_t = D \left(\frac{Z_t}{\hat{F}_{t-s}} \right) + (1 - D) \hat{F}_{t-s}, 0 < D < 1, t = s + 1, \dots, N, \quad (2.21)$$

$$\bar{Z}_t = A \left(\frac{Z_t}{\hat{F}_{t-s}} \right) + (1 - A) (\hat{Z}_{t-1} + \hat{T}_{t-1}), 0 < A < 1, t = s + 1, \dots, N, \quad (2.22)$$

$$\hat{T}_t = C (\bar{Z}_t - \bar{Z}_{t-1}) + (1 - C) \hat{T}_{t-1}, 0 < C < 1, t = s + 1, \dots, N, \quad (2.23)$$

em que A , C e D , representam as constantes de suavização.

No caso de sazonalidade aditiva, ou seja, o padrão sazonal não se altera com a mudança de nível, a equação geral do modelo é dada por:

$$Z_t = \mu_t + T_t + \varepsilon_t. \quad (2.24)$$

As equações de suavização são modificadas para:

$$\hat{F}_t = D(Z_t - \bar{Z}_t) + (1 - D)\hat{F}_{t-s}, 0 < D < 1, \quad (2.25)$$

$$\bar{Z}_t = A(Z_t - \hat{F}_{t-s}) + (1 - A)(\bar{Z}_{t-1} + \hat{T}_{t-1}), 0 < A < 1, \quad (2.26)$$

$$\hat{T}_t = C(\bar{Z}_t - Z_{t-1}) + (1 - C)\hat{T}_{t-1}, 0 < C < 1, \quad (2.27)$$

em que A , C e D , representam as constantes de suavização.

A previsão é dada por:

$$\hat{Z}_t(h) = (\bar{Z}_t + h\hat{T})_t \hat{F}_{t+h-s}, h = 1, 2, \dots, s. \quad (2.28)$$

A determinação das constantes de suavização é feita de maneira que a soma dos quadrados dos erros ajustados seja mínima (MORETTIN e TOLOI, 2004).

Segundo Morettin e Toloi (2004), as vantagens deste método são semelhantes às vantagens do método de Holt e aplicam-se a séries que apresentam comportamento geral. As desvantagens estão nas dificuldades em determinar os valores mais adequados das constantes de suavização e na impossibilidade de se estudar as suas propriedades estatísticas e de construir um intervalo de confiança.

2.2.2 Modelos ARIMA

A classe dos modelos ARIMA (Autorregressivos Integrados de Médias Móveis) ou Modelos de Box e Jenkins, é composta por modelos que visam captar o comportamento da autocorrelação entre os valores da série temporal e com base nesse comportamento é possível realizar previsões (TÍBULO, 2014). Esta classe de modelos é capaz de descrever satisfatoriamente séries estacionárias e séries não estacionárias, desde que não apresentem comportamento explosivo (MORETTIN e TOLOI, 2004).

A construção do modelo será baseada em um ciclo iterativo no qual a escolha da estrutura do modelo é baseada nos próprios dados. As fases desse ciclo são:

1. Identificação: Identifica-se um modelo adequado para a série em questão com base na análise de autocorrelações, autocorrelações parciais e outros critérios;

2. Estimação: Os parâmetros do modelo identificado são estimados;
3. Verificação ou Diagnóstico: É necessário checar a adequação do modelo por meio da análise de resíduos;
4. Previsão: O modelo final é utilizado na previsão de valores futuros para a série.

Com a verificação, caso o modelo não seja adequado, repete-se o ciclo, até que o modelo ideal seja escolhido. Segundo Morettin e Toloi (2004), a fase mais crítica desse ciclo iterativo é a fase da identificação. O autor afirma ainda a possibilidade de que vários pesquisadores identifiquem modelos diferentes para a mesma série temporal.

De acordo com Morettin e Toloi (2004), os modelos postulados são parcimoniosos, uma vez que contêm um número pequeno de parâmetros e as previsões obtidas são bastante precisas, comparando-se com os demais métodos de previsão.

2.2.2.1 Modelos Autorregressivos (AR)

Os modelos autorregressivos são denotados por $AR(p)$, sendo p o número de defasagens de Z_t . A observação Z , no período t é gerada por uma média ponderada das observações passadas, sendo que p períodos anteriores são incluídos na série, juntamente com o ruído.

Um modelo $AR(p)$ pode ser representado por:

$$\bar{Z}_t = \phi_1 \bar{Z}_{t-1} + \phi_2 \bar{Z}_{t-2} + \dots + \phi_p \bar{Z}_{t-p} + \varepsilon_t. \quad (2.29)$$

Utilizando o operador autorregressivo B :

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p, \quad (2.30)$$

pode-se escrever:

$$\phi(B) \bar{Z}_t = \varepsilon_t. \quad (2.31)$$

Os coeficientes autorregressivos (ϕ_p) são parâmetros que descrevem como um valor atual se relaciona com seus valores passados e podem ser estimados através do método dos mínimos quadrados. O número de defasagens p a ser utilizado no modelo é resultante do exame dos valores da função de autocorrelação parcial (FACP).

Cada ε_t é considerado como uma variável aleatória de média zero, variância constante e covariância igual a zero para $k \neq 0$.

Quanto à condição de estacionariedade, segundo Morettin e Toloi (2004), um modelo AR será estacionário se todas as raízes $\phi(B) = 0$ do operador AR forem maiores que 1, isto é, todas as raízes devem estar fora do círculo unitário.

2.2.2.2 Modelos de Médias Móveis (MA)

Os modelos de Médias Móveis $MA(q)$ são gerados por meio de uma média ponderada dos erros aleatórios de defasagem q . São semelhantes aos modelos AR , porém, em vez de considerar os valores passados da série, neste modelo, considera-se os valores correntes dos erros aleatórios e seus q valores passados. A equação do modelo $MA(q)$ é dada abaixo:

$$Z_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}. \quad (2.32)$$

Utilizando-se o operador de médias móveis:

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q, \quad (2.33)$$

tem-se que:

$$\theta(B)\varepsilon_t = Z_t. \quad (2.34)$$

Os parâmetros (θ_q) podem ser tanto positivos quanto negativos. O número de defasagens q a ser utilizado no modelo é resultante do exame dos valores da função de autocorrelação (FAC).

Cada ε_t é considerada como uma variável aleatória de média zero, variância constante e covariância igual a zero para $k \neq 0$.

Desde que $\psi(B) = 1 - \theta_1 B - \dots - \theta_q B^q$, não há restrições sobre os parâmetros θ_j para que o processo seja estacionário.

2.2.2.3 Modelos Autorregressivos e de Médias Móveis (ARMA)

Os modelos Autorregressivos e de Médias Móveis $ARMA(p,q)$ resultam da combinação de processos Autorregressivos $AR(p)$ e de processos Médias Móveis $MA(q)$.

Tais modelos relacionam os valores futuros com os valores observados no passado, sendo especificados por:

$$\tilde{Z}_t = \phi_1 \tilde{Z}_{t-1} + \dots + \phi_p \tilde{Z}_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q}. \quad (2.35)$$

Sendo $\phi(B)$ e $\theta(B)$, respectivamente, operadores autoregressivos e de médias móveis, a equação acima pode ser reescrita como:

$$\phi(B)\tilde{Z}_t = \theta(B)\varepsilon_t. \quad (2.36)$$

Um processo $ARMA(p, q)$ será estacionário se as raízes de $\phi(B) = 0$ caírem todas fora do círculo unitário (MORETTIN e TOLOI, 2004). Ou seja, a condição de estacionariedade para um modelo $ARMA(p, q)$ é a mesma que para um processo $AR(p)$.

Os três modelos apresentados até aqui ($AR(p)$, $MA(q)$ e $ARMA(p, q)$) possuem em comum o fato de considerarem a série em análise como estacionária, isto é, $d = 0$.

2.2.2.4 Modelos Autorregressivos Integrados de Médias Móveis (ARIMA)

Os modelos apresentados anteriormente não são apropriados para descrever séries não estacionárias. Morettin e Toloi (2004), afirmam que, na prática, muitas séries não são estacionárias, porém quando diferenciadas tornam-se estacionárias. É aí que entram os modelos Autorregressivos Integrados de Médias Móveis, $ARIMA(p, d, q)$, que se tratam de modelos diferenciados d vezes na série original, possuem p parâmetros autoregressivos e q parâmetros de médias móveis. Desta forma d representa a ordem de integração, isto é, o número de diferenças necessárias para que a série não-estacionária satisfaça a condição de estacionariedade. Um processo $ARIMA(p, d, q)$, portanto, nada mais é do que um processo $ARMA(p, q)$ integrado.

A equação do modelo ARIMA é dada abaixo:

$$\Delta^d Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \dots + \phi_p Z_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}, \quad (2.37)$$

em que os parâmetros ϕ_1, \dots, ϕ_p são referentes à parte autoregressiva e $\theta_1, \dots, \theta_q$ são parâmetros de médias móveis. ε_t é o erro, o qual não pode ser estimado pelo modelo e d representa a ordem de integração utilizada no modelo.

O modelo pode ainda ser descrito reduzidamente por:

$$\phi(B)\Delta^d Z_t = \theta(B)\varepsilon_t. \quad (2.38)$$

Geralmente, apenas uma ou duas diferenciações são necessárias na série para que esta se torne estacionária. Estes modelos, segundo Morettin e Toloi (2004), são apropriados para representar séries cujo comportamento seja não explosivo em seus dados.

2.2.2.5 Modelos Autorregressivos Integrados de Médias Móveis Sazonal (SARIMA)

Grande parte das séries temporais apresenta variações sazonais, isto é, exibem um comportamento periódico que se repete a cada s intervalos de tempo, sendo necessário, portanto, a utilização de modelos que incorporem tal característica. Neste cenário os modelos $ARIMA(p, d, q)$ foram ampliados para os modelos $SARIMA(p, d, q)(P, D, Q)_{12}$, os quais podem ser representados pela equação:

$$\phi(B)\Phi(B^{12})(1 - B^{12})^D(1 - B)^d Z_t = \theta(B)\Theta(B^{12})\alpha_t, \quad (2.39)$$

em que:

$$\theta(B) = 1 - \theta_1 B - \dots - \theta_q B^q;$$

$$\phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p;$$

$$\Phi(B^{12}) = 1 - \Phi_1 B^{12} - \dots - \Phi_P B^{12P};$$

$$\Theta(B^{12}) = 1 - \Theta_1 B^{12} - \dots - \Theta_Q B^{12Q};$$

$$(1 - B)^d = \Delta^d;$$

$$(1 - B^{12})^D = \Delta_{12}^D$$

e $\theta(B)$ é o operador de médias móveis simples de ordem q , $\phi(B)$ é o operador autorregressivo simples de ordem p , $\Theta(B^{12})$ é o operador de médias móveis sazonal de ordem Q , $\Phi(B^{12})$ é o operador autorregressivo sazonal de ordem P , Δ^d é o operador diferença simples com d indicando o número de diferenças simples e Δ_{12}^D é o operador diferença sazonal com D indicando o número de diferenças sazonais.

2.3 PROCEDIMENTOS ÚTEIS NA MODELAGEM DE SÉRIES TEMPORAIS

Segundo Morettin e Tolo (2004), um primeiro passo na análise de uma série temporal é a construção de seu gráfico, uma vez que, tal procedimento revelará características importantes como tendências, sazonalidade, variabilidade, observações atípicas, etc. Ainda de acordo com os autores, é aconselhável estabelecer se existe componente de tendência na série antes de aplicar qualquer procedimento para sua estimação. Portanto, além dessa inspeção gráfica, é possível e necessário fazer uso de testes de hipóteses estatísticos para verificar se a série é estacionária e se existe tendência e/ou sazonalidade na mesma.

2.3.1 Testes de Raiz Unitária Simples

2.3.1.1 Teste de Dickey-Fuller

O objetivo do teste de Dickey-Fuller (DF) é detectar a presença de raiz unitária nos polinômios autorregressivos das séries temporais. O número de raízes unitárias ou ordem de integração é dado pelo número de vezes que a série foi diferenciada para se tornar estacionária. Segundo Morettin e Tolo (2004), quando a série temporal é estacionária, as propriedades estatísticas (média, variância e autocorrelação) observadas no passado serão as mesmas no futuro.

Basicamente, o teste de Dickey-Fuller estima a seguinte autorregressão:

$$\nabla Z_t = \gamma Z_{t-1} + \varepsilon_t, \quad (2.40)$$

isto é:

$$\nabla Z_t = (\rho - 1)Z_{t-1} + \varepsilon_t, \quad (2.41)$$

em que $\nabla = (Z_t - Z_{t-1})$ e $\gamma = \rho - 1$. Dessa forma, a hipótese nula H_0 é de que exista pelo menos uma raiz unitária de maneira que a série não é estacionária e $\gamma = 0$. Por outro lado, a hipótese alternativa H_1 é que a variável é estacionária, não possuindo nenhuma raiz unitária e consequentemente $\gamma < 0$.

O teste Dickey-Fuller parte da pressuposição de que os termos de erros da série temporal em questão são independentes e identicamente distribuídos, ou seja, não apresentam correlação. Dessa maneira, o teste Dickey-Fuller foi ampliado de forma a incorporar

defasagens em relação à variável em análise. Assim procedendo, obtém-se o teste Dickey-Fuller Aumentado (ADF). As hipóteses do teste ADF, de acordo com Morettin e Tolo (2004), são:

$H_0: \phi_1 = 1$: a série possui raiz unitária, ou seja, é não estacionária;

$H_1: \phi_1 < 1$: a série não possui raiz unitária, ou seja, é estacionária.

No teste de Dickey-Fuller Aumentado, a estatística de teste é dada por:

$$\hat{\alpha} = \frac{\sum_{t=1}^N Z_{t-1} Z_t}{\sum_{t=1}^N Z_{t-1}^2} - 1. \quad (2.42)$$

Estima-se em seguida a variância amostral:

$$S^2 = \frac{1}{N-1} \sum_{t=1}^N (\Delta Z_t - \hat{\alpha} Z_{t-1})^2. \quad (2.43)$$

Calcula-se o desvio padrão do coeficiente:

$$S(\hat{\alpha}) = \frac{S}{\sqrt{\sum_{t=1}^T Z_{t-1}^2}}. \quad (2.44)$$

O valor calculado da estatística é dado por:

$$\hat{\tau} = \frac{\hat{\alpha}}{S(\hat{\alpha})}. \quad (2.45)$$

Os valores críticos da estatística τ foram tabelados por Dickey e Fuller através de Simulação Monte Carlo. A hipótese nula é rejeitada se $\hat{\tau} < \tau$.

De acordo com Margarido e Junior (2006), a principal vantagem do teste de Dickey-Fuller Aumentado em relação ao teste de Dickey-Fuller é que ao introduzir um número suficiente de defasagens, garante-se que os resíduos não apresentem correlação.

Uma desvantagem do teste, no entanto, é seu baixo poder estatístico, há baixa probabilidade de se rejeitar a hipótese nula quando esta é falsa, o que, posto de maneira simples, torna o teste praticamente incapaz de diferenciar uma série com raiz unitária de uma com raiz “quase” unitária, criando um viés para conclusão da existência de raiz unitária.

2.3.1.2 Teste de Phillips-Perron

O teste de Phillips-Perron (PP) é uma generalização do teste de Dickey-Fuller para os casos em que os erros ε_t são correlacionados e possivelmente heterocedásticos. As estatísticas consideradas por Phillips e Perron são modificadas para levar em conta a autocorrelação, após estimar os modelos por Mínimos Quadrados Ordinários (MORETTIN e TOLOI, 2004).

A estatística de teste é dada por:

$$\hat{Z}_{t,\mu} = \frac{\hat{\sigma}\hat{\tau}_\mu}{\hat{\lambda}} - \frac{1}{2} \frac{(\hat{\lambda}^2 - \hat{\sigma}^2)}{\hat{\lambda}\sqrt{T^{-2} \sum_{t=1}^T y_{t-1}^2}}, \quad (2.46)$$

em que, teoricamente, $\hat{\lambda}^2$ é a variância de longo prazo de ε_t (limite de T vezes a variância da média amostral de μ_t), $\hat{\tau}_\mu$ é a estatística DF e $\hat{\sigma}^2$ é um estimador consistente da variância populacional da regressão. As hipóteses para o teste são:

H_0 : A série é não estacionária, isto é, possui raiz unitária;

H_1 : A série é estacionária, isto é, não possui raiz unitária.

A intuição dessa modificação feita por Phillips e Perron é que, na presença de autocorrelação serial, é necessário levar em conta não somente a variância dos resíduos como suas autocovariâncias. O importante desse termo de correção é que este é uma função de diferença entre a variância estimada de μ_t e a variância de longo prazo estimada de μ_t . Assim o termo de correção zera já que $\lambda^2 = \sigma^2$. Ao mesmo tempo, esse termo de correção impede que a distribuição assintoticase altere na presença de correlação serial (BRAGA, 2008).

Assim como o teste de Dickey-Fuller, o teste de Phillips-Perron possui baixo poder estatístico, isto é, há baixa probabilidade de se rejeitar a hipótese nula quando esta é falsa.

2.3.1.3 Teste de Kwiatkowski, Philips, Schmidt e Shin

O teste Kwiatkowski-Phillips-Schmidt-Shin (KPSS) foi criado como contraponto aos tradicionais testes de raízes unitárias em que a hipótese nula é de existência de raiz unitária. Neste teste a hipótese nula é a de estacionariedade da série. Dessa forma pode-se lançar mão do teste Kwiatkowski-Philips-Schmidt-Shin conjuntamente aos testes de Dickey-Fuller e de Phillips-Perron a fim de corroborar a presença de raiz unitária ou não na série.

A ideia é testar a variância de passeio aleatório. Se essa variância for nula, então o processo é estacionário (BUENO, 2008 *apud* KLIDZIO, 2009). A estatística para o teste KPSS é dada pela expressão

As hipóteses do teste são:

H_0 : A série é estacionária, isto é, não possui raiz unitária;

H_1 : A série é não estacionária, isto é, possui raiz unitária.

A estatística de teste é dada por:

$$LM_{KPSS} = \sum_{t=1}^T \frac{s_t^2}{T^2 \hat{\lambda}^2}, \quad (2.47)$$

em que $\hat{\lambda}^2$ representa a variância de longo prazo de ε_t , semelhante ao teste PP.

Caso a estatística calculada seja menor que os valores críticos tabulados por KPSS (1992), H_0 não é rejeitada, ou seja, a série apresenta estacionariedade.

2.3.2 Testes para Tendência

2.3.2.1 Teste de Wald-Wolfowitz

Seja $Z_t, t = 1, \dots, N$ uma série temporal com N observações. Seja m a mediana desses valores. Atribui-se a cada valor Z_t o símbolo A , caso ele seja maior ou igual a m , e B , caso ele seja menor que m . Têm-se então $N = (n_1 \text{ pontos } A) + (n_2 \text{ pontos } B)$. A estatística usada no teste é:

T_1 = número total de sequências (grupos com símbolos iguais).

As hipóteses consideradas são:

H_0 : Não há tendência;

H_1 : Há tendência.

Rejeita-se a hipótese nula H_0 se há poucas sequências, isto é, se T_1 for relativamente pequeno. Para um dado nível de significância α , rejeita-se H_0 se $T_1 < W_\alpha$, onde W_α é o α -quantil da distribuição de T_1 , o qual é tabelado.

Para n_1 ou n_2 maior que 20, pode-se usar a aproximação normal, dada por: $T_1 \sim N(\mu, \sigma^2)$, onde:

$$\mu = \frac{2n_1n_2}{N} + 1, \quad (2.48)$$

$$\sigma = \sqrt{\frac{2n_1n_2(2n_1n_2 - N)}{N^2(N - 1)}}. \quad (2.49)$$

2.3.2.2 Teste de Cox-Stuart

Considerando-se um conjunto de observações Z_1, Z_2, \dots, Z_N , agrupa-se as observações em pares $(Z_1, Z_{1+c}), (Z_2, Z_{2+c}), \dots, (Z_N, Z_{N-c}, Z_N)$, onde $c = \frac{N}{2}$, se N for par e $c = \frac{N+1}{2}$, se N for ímpar. A cada par de observações é associado o sinal positivo ‘+’ se $Z_i < Z_{i+c}$ e o sinal negativo ‘-’ se $Z_i > Z_{i+c}$. Caso $Z_i = Z_{i+c}$, elimina-se esta observação. Seja n o número de pares onde $Z_i \neq Z_{i+c}$, procura-se testar:

$H_0: P(Z_i < Z_{i+c}) = P(Z_i > Z_{i+c}), \forall i$: Não existe tendência;

$H_1: P(Z_i < Z_{i+c}) \neq P(Z_i > Z_{i+c}), \forall i$: Existe tendência.

Este é um teste bilateral, no qual a estatística de teste é dada por:

T = Número de pares com sinal positivo ‘+’.

Para uma amostra com número de observações $n \leq 20$, faz-se uso da distribuição binomial para avaliar o teste com os parâmetros $p = 0,5$ e $n = N_t$ com hipóteses do teste bilateral e para uma amostra com número de observações $n > 20$ pode-se usar a aproximação normal.

2.3.2.3 Teste Baseado no Coeficiente de Correlação de Spearman

Segundo Morettin e Toloi (2004), este teste é mais poderoso que o teste de Cox-Stuart. Denomine-se R_t o posto de Z_t dentre as N observações. Caso haja observações empatadas, devem-se usar os postos médios.

$$\text{Seja } T_3 = \sum_{t=1}^N [R_t - t]^2, \quad (2.50)$$

em que $t = 1, 2, \dots$ são postos ‘naturais’ dos instantes de tempo. Rejeita-se a hipótese nula H_0 de que não existe tendência se T_3 for grande ou pequeno.

A distribuição de T_3 é tabelada pela tabela de Conover (1980) (MORETTIN e TOLOI, 2004). O teste também pode ser realizado em função do coeficiente de correlação de Spearman, dado por:

$$\rho = 1 - \frac{6T_3}{N(N^2 - 1)}. \quad (2.51)$$

2.3.2.4 Teste de Mann-Kendall

Segundo Salviano *et al* (2016), o teste de Mann-Kendall é utilizado para determinar se determinada série de dados possui uma tendência temporal de alteração estatisticamente significativa. Como é um teste não paramétrico, não requer a distribuição normal dos dados, além de ser pouco influenciado por mudanças abruptas ou séries não homogêneas. Porém, tem por exigência que os dados sejam independentes e aleatórios.

Sejam Z_1, Z_2, \dots, Z_n observações de uma série temporal. Procura-se testar se as observações da série são independentes e identicamente distribuídas, isto é, procura-se testar as seguintes hipóteses:

H_0 : As observações da série são independentes e identicamente distribuídos (não há tendência);

H_1 : As observações da série possuem tendência no tempo (há tendência).

A estatística de teste, sob H_0 , é dada por:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sgn}(x_j - x_i), \quad (2.52)$$

$$\text{em que } \text{sgn}(x_j - x_i) = \begin{cases} +1; & \text{se } x_j > x_i \\ 0; & \text{se } x_j = x_i \\ -1; & \text{se } x_j < x_i \end{cases}$$

Quando $n \geq 10$, S pode ser comparada a uma distribuição normal com média $\mu = 0$ e variância dada por:

$$\sigma^2 = \frac{n(n-1)(2n+5) - \sum_{i=1}^n t_i * (i)(i-1)(2i+5)}{18}. \quad (2.53)$$

A estatística de teste Z segue uma distribuição normal, com média igual a zero, em que valores positivos indicam uma tendência crescente e valores negativos indicam tendência decrescente. Com base no sinal de S , a estatística de teste pode ser calculada por:

$$Z = \begin{cases} \frac{S-1}{\sigma}, & \text{se } S > 0 \\ 0, & \text{se } S = 0 \\ \frac{S+1}{\sigma}, & \text{se } S < 0. \end{cases} \quad (2.54)$$

Como se trata de um teste bicaudal, a hipótese nula H_0 será rejeitada se o valor absoluto de Z for superior a $\frac{Z_{\alpha}}{2}$.

2.3.3 Testes de Raiz Unitária Sazonal

2.3.3.1 Teste HEGY

O teste HEGY é uma extensão do teste de Dickey-Fuller na frequência zero. No teste, se especifica um modelo autorregressivo para a série Z_t ,

$$\varphi(B)Z_t = \varepsilon_t, \quad (2.55)$$

em que $\varphi(B)$ é um polinômio no operador de retardos B , onde cada uma de suas raízes corresponde a flutuações em cada uma das frequências sazonais e ε_t é um processo não correlacionado no tempo. O objetivo do teste é conhecer se as raízes do polinômio $\varphi(B)$ tem módulo unitário ou não. No caso de séries trimestrais, as estatísticas de teste de constroem a partir da seguinte regressão auxiliar (RODRIGUÉZ *et al*, 2012):

$$\varphi(B)^* y_{4t} = \pi_1 y_{1,t-1} + \pi_2 y_{2,t-1} + \pi_3 y_{3,t-2} + \pi_4 y_{3,t-1} + \varepsilon_t, \quad (2.56)$$

em que cada uma das variáveis $y_{k,t}$ possui apenas uma das frequências sazonais efetivamente isoladas das demais. A regressão é estimada através de mínimos quadrados ordinários possivelmente incluindo retardos adicionais de y_{4t} de modo a tornar os erros ruídos brancos. O teste se estabelece a partir das estatísticas T e F , comparando-as com valores críticos aproximados a partir de simulação de Monte Carlo. Para as frequências 0 e π se utilizam as estatísticas T , enquanto a frequência $\frac{\pi}{2}$ é posta à prova utilizando-se uma estatística F . Desta

maneira, a hipótese nula corresponde à existência de raízes unitárias nas frequências sazonais e na frequência zero, enquanto a hipótese alternativa implica na existência de sazonalidade estacionária.

2.3.3.2 Teste de Osborn-Chui-Smith-Birchenhall

O teste de Osborn-Chui-Smith-Birchenhall é habitualmente designado teste de OCSB e é o teste mais comum para se obter o número de diferenciações sazonais necessárias para tornar uma série estacionária na parte sazonal (HYNDMAN, 2012 *apud* BARBOSA, 2014).

A regressão do teste tem a forma:

$$\Delta\Delta_{12}Z_t = \beta_1\Delta_{12}Z_{t-1} + \sum_{k=1}^{12} \delta_k D_{kt} + \sum_{i=1}^p \phi_i \Delta\Delta_{12}Z_{t-i} + \varepsilon_t, \quad (2.57)$$

em que, D_k é uma variável sazonal que toma o valor 1 no mês k e 0 no restante dos meses e é um termo de erro.

A equação do teste para uma série mensal é dada por:

$$\Delta\Delta_{12}Z_t = \sum_{k=1}^{12} \delta_k D_{kt} + \beta_1\Delta_{12}Z_{t-1} + \beta_2Z_{t-12} + \sum_{j=1}^p \gamma_j \Delta\Delta_{12}Z_{t-j} + \varepsilon_t. \quad (2.58)$$

A ordem p da equação anterior para o modelo é seleccionada mediante o Critério de Akaike (ver seção 4.3.4.1), escolhendo-se aquele que apresentar o menor valor para este critério.

H_0 : A série é não estacionária na parte sazonal;

H_1 : A série é estacionária na parte sazonal.

Para um nível de significância $< \alpha$, a hipótese nula é rejeitada a um nível α de significância.

2.3.3.3 Teste de Canova-Hansen

Outro teste que analisa a existência de raiz unitária sazonal é o teste CH, proposto por Canova e Hansen. O teste é aplicado para verificar se a mudança no padrão sazonal ao longo

do tempo é suficiente para justificar uma raiz unitária sazonal. Neste teste a hipótese nula considera a estacionariedade sazonal da série, ficando a hipótese de existência de raiz unitária sazonal para a hipótese alternativa. Este teste é uma extensão do teste KPSS (ZAMPROGNO, 2004).

O teste pode ser visto como um contraste de raiz unitária e como uma forma de contrastar a instabilidade no padrão sazonal (RODRIGUÉZ *et al*, 2012). O mesmo propõe uma regressão sobre os indicadores sazonais e outros componentes deterministas, uma vez que, considera os resíduos de uma regressão onde o regressando é a variável em níveis ou sua primeira diferença se a série for integrada de ordem um (possui uma unidade de raiz na frequência zero) e os regressores são componentes deterministas no modelo especificado na hipótese nula. Neste caso, os níveis para cada estação representam o componente determinista da sazonalidade e são assumidos como permanentes constantes para toda a amostra. A regressão considerada por Canova e Hansen é a seguinte:

$$y_t = \mu + x_t'\beta + f_t'\gamma + \varepsilon_t, \quad (2.59)$$

em que y_t é a série a que se aplica o contraste $y_t = 1, \dots, T$ (T é o número de observações), μ é uma constante, x_t contém regressores não estocásticos e f_t é um processo sazonal determinístico de média zero.

De acordo com Canova e Hansen, uma forma razoável de modelar a sazonalidade evolutiva é considerar que os coeficientes γ variem através do tempo, como por exemplo, um passeio aleatório.

$$\gamma_t = \gamma_{t-1} + \mu_t. \quad (2.60)$$

Quando a matriz de variâncias e covariâncias de μ_t for idêntica a zero, a sazonalidade é do tipo determinística. A hipótese nula surge a partir de uma generalização das linhas anteriores, portanto, a não significância do parâmetro estimado implica que a hipótese nula não pode ser rejeitada, classificando a sazonalidade como determinística. A existência de raízes unitárias nas frequências semestrais e anuais é contrastada. No artigo original de Canova e Hansen (1995), os autores oferecem duas alternativas para o processo, funções sazonais e trigonométricas.

Uma vez que o modelo inicial é estabelecido, o mesmo é estimado por meio de mínimos quadrados ordinários. A partir dos resíduos correspondentes a cada estação (e suas somas parciais), se contrasta a seguinte estatística de teste:

$$L = \frac{\sum_{t=1}^T \hat{F}_t' A (A' \hat{\Omega}^f A)^{-1} A' \hat{F}_t}{T^2}. \quad (2.61)$$

A estatística de teste é comparada com valores críticos (assintóticos) correspondentes a distribuição de Von Mises.

2.3.4 Testes para Sazonalidade

2.3.4.1 Teste de Kruskal-Wallis

Morettin e Toloi (2004), apresentam o teste de Kruskal-Wallis como alternativa para identificar sazonalidade em séries temporais. Assumindo que os dados são coletados em períodos de tempo equiespaçados, cada coluna K é considerada uma amostra de tamanho n_j . Dessa maneira as observações são:

$$Z_{ij}, j = 1, \dots, k, i = 1, \dots, n_j, N = \sum_{j=1}^k n_j. \quad (2.62)$$

As observações de Z_{ij} são substituídas por seus postos R_{ij} , obtidos ordenando-se todas as N observações. Seja R_j a soma dos postos associados à j -ésima amostra:

$$R_j = \sum_{i=1}^{n_j} R_{ij}, j = 1, \dots, k. \quad (2.63)$$

A estatística de teste é dada por:

$$T_1 = \frac{12}{N(N+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} - 3(N+1). \quad (2.64)$$

As hipóteses consideradas são:

H_0 : Não há sazonalidade;

H_1 : Há sazonalidade.

A hipótese nula H_0 é rejeitada se a estatística de teste for maior ou igual ao valor crítico T_{1c} (valor tabelado da χ^2) ($T_1 \geq T_{1c}$), para dado nível de significância. O valor crítico

T_{1c} é tal que $P_h(T_1 \geq T_{1c}) = \alpha$, em que α é o nível de significância do teste. Para n_j suficientemente grande, ou $k \geq 4$, sob H_0 , a distribuição T_1 pode ser aproximada por uma variável χ^2 com $k - 1$ graus de liberdade.

2.3.4.2 Teste de Friedman

O teste de Friedman divide a série temporal em blocos de períodos e calcula o posto de cada observação em cada bloco, os quais são considerados independentes.

As hipóteses consideradas são:

H_0 : Não há sazonalidade;

H_1 : Há sazonalidade.

De acordo com Morettin e Toloi (2004), a estatística do teste de Friedman é dada por:

$$T_2 = \frac{12}{pk(k+1)} \sum_{j=1}^k R_{.j}^{*2} - 3p(k+1), \quad (2.65)$$

em que p = número de blocos = número de anos, k = número de tratamentos = 12 e $R_{.j}^*$ denota a soma dos postos da j -ésima coluna, isto é,

$$R_{.j}^* = \sum_{i=1}^p R_{ij}^*, \quad (2.66)$$

em que R_{ij}^* = posto de Z_{ij} dentro do bloco i , de 1 até k . A distribuição de T_2 pode ser aproximada por um χ^2 com $k - 1$ grau de liberdade.

2.3.4.3 Teste de Fisher

O teste de Fisher foi proposto por Priestley (1989) para verificar a existência de periodicidade, também denominada sazonalidade em séries temporais. Esse teste utiliza os valores do periodograma, sendo capaz de detectar grandes periodicidades. O teste é baseado nas seguintes hipóteses:

H_0 : Não há periodicidade;

H_1 : Há periodicidade.

A estatística de teste é dada por:

$$g = \frac{\max I_j^{(n)}}{\sum_{j=1}^{\frac{n}{2}} I_j^{(n)}}, \quad (2.67)$$

em que I_j é o valor do periodograma no período j e n é o tamanho da amostra.

A estatística Z é dada por:

$$Z_\alpha = 1 - \left(\frac{\alpha}{\frac{n}{2}} \right)^{\frac{1}{\frac{n}{2}-1}}, \quad (2.68)$$

em que α é o nível de significância do teste.

Para $g > Z_\alpha$, H_0 é rejeitada, isto é, a série possui sazonalidade no período j .

2.3.5 Transformações

Após aplicação de testes estatísticos e constatação de tendência e/ou sazonalidade na série temporal em estudo, será necessário transformar os dados originais. Isto porque, como já mencionado anteriormente, a maioria dos procedimentos de análise estatística de séries temporais supões que estas sejam estacionárias (MORETTIN e TOLOI, 2004).

Segundo Morettin e Toloí (2004), os dados originais podem ser transformados ainda com o objetivo de se estabilizar a variância. Caso a série não apresente variância constante, esta deverá ser transformada a fim de estabilizá-la. Outro objetivo do uso de transformações é se obter uma distribuição para os dados mais simétrica e próxima da distribuição normal. No entanto, os autores citam Nelson (1976), o qual conclui que transformações não melhoram a qualidade da previsão; Markridakis e Hibon (1979), os quais verificaram que os dados transformados têm pouco efeito na melhoria da previsão; e Granger e Newbold, que concluíram que depois de transformados os dados, é introduzido um viés nas previsões, decorrente de tal transformação, além disso, os mesmos observaram que a heterocedasticidade não afeta a adequação da previsão, uma vez que, não implicam em estimadores viesados. Diante do exposto, transformações de dados devem ser usadas cuidadosamente.

2.3.5.1 Diferenciação

De acordo com Morettin e Tolo (2004), a transformação mais comum consiste em tomar diferenças sucessivas da série original, até se obter uma série estacionária. Comumente, são necessárias uma ou duas diferenças para que a série se torne estacionária.

A primeira diferença de Z_t é definida por:

$$\Delta Z_t = Z(t) - Z(t - 1). \quad (2.69)$$

A segunda diferença é:

$$\Delta^2 Z(t) = \Delta[\Delta Z_t] = \Delta[Z(t) - Z(t - 1)], \quad (2.70)$$

isto é,

$$\Delta^2 Z(t) = Z(t) - 2Z(t - 1) + Z(t - 2). \quad (2.71)$$

Desta forma, a n -ésima diferença de $Z(t)$ pode ser definida por:

$$\Delta^n Z(t) = \Delta[\Delta^{n-1} Z(t)]. \quad (2.72)$$

Analogamente, pode-se remover a sazonalidade tomando-se diferenças na defasagem sazonal. Nestes casos, o número de termos sazonais é raramente mais do que um.

Dado o período sazonal S , a diferenciação sazonal é dada por:

$$\Delta_S^D Z(t) = Z(t) - Z(t - S). \quad (2.73)$$

2.3.5.2 Transformação Box-Cox

Para séries heterocedásticas (que não apresentam variância constante), é necessário transformar a série original para estabilizá-la. De acordo com Morettin e Tolo (2004), a heterocedasticidade é comum em séries econômicas e financeiras, requerendo alguma transformação não-linear, como alguma transformação de Box-Cox. Estas obtêm uma série mais simétrica e com variância constante, isto é, mais próxima da distribuição gaussiana (distribuição normal) quando valores apropriados de λ na equação abaixo são encontrados:

$$Z_t^{(\lambda)} = \begin{cases} \frac{Z_t^\lambda - c}{\lambda}, & \text{se } \lambda \neq 0 \\ \log Z_t, & \text{se } \lambda = 0. \end{cases} \quad (2.74)$$

O valor de λ e c são parâmetros a serem estimados. O λ é obtido a partir de uma varredura sobre um intervalo de valores possíveis, na qual se procura identificar aquele que apresente a menor soma dos quadrados dos resíduos.

Um fator importante a ser considerado é que quando uma diferenciação simples ou sazonal estiver em uso simultâneo com uma transformação de Box-Cox, a transformação que busca estabilizar a variância sempre deve ser aplicada primeiramente.

2.4 METODOLOGIA BOX E JENKINS

Há variadas aplicações gerais e específicas empreendidas por diversos autores quanto à metodologia Box e Jenkins. Tais modelos possuem aplicações na área médica, ambiental, financeira e de engenharia, modelagem de séries temporais para a previsão da qualidade do ar, modelagem da demanda de turistas, etc. (WERNER; RIBEIRO, 2003 *apud* JACOBS, 2011).

Os modelos Box e Jenkins partem do princípio de que cada valor da série temporal pode ser explicado por seus valores prévios, em virtude da correlação temporal que geralmente há entre os valores da série. Pressupondo-se que os valores de determinada série temporal são altamente dependentes dos valores passados, seus valores futuros poderão ser explicados pelos valores passados da série (PELLEGRINI, 2000 *apud* JACOBS, 2011).

A metodologia consiste, portanto, em ajustar modelos autorregressivos integrados de médias móveis, *ARIMA* (p, d, q), a um conjunto de dados (MORETTIN e TOLOI, 2004).

Conforme já visto na subseção 2.2.2, esta permite que valores futuros de uma série sejam previstos tomando-se por base apenas seus valores presentes e passados e está fundamentada em um ciclo iterativo de quatro estágios:

1. Identificação: Identifica-se um modelo adequado para a série em questão com base na análise de autocorrelações, autocorrelações parciais e outros critérios;
2. Estimação: Os parâmetros do modelo identificado são estimados;
3. Verificação ou Diagnóstico: É necessário checar a adequação do modelo por meio da análise de resíduos;
4. Previsão: O modelo final é utilizado na previsão de valores futuros para a série.

Caso o modelo encontrado não seja adequado, o ciclo é repetido, voltando-se à fase de identificação.

2.4.1 Identificação

Jacobs (2011) define a etapa da identificação como a mais importante do processo, pois caso seja adotado um modelo inadequado, os resultados obtidos por meio deste podem ser desastrosos.

Segundo Morettin e Toloi (2004), um procedimento que muitas vezes é adotado é identificar nesta fase mais de um modelo a ser estimado e verificado. Dessa maneira, pode-se comparar posteriormente o erro quadrático médio de previsão entre eles e, finalmente, escolher, dentre os modelos ajustados, o melhor.

Na fase da identificação, devem ser determinadas as ordens simples (p, q) e/ou sazonais (P, Q) do modelo, assim como as ordens de integração (d, D) , além de estimativas preliminares dos parâmetros a serem usados no estágio da estimação (MORETTIN e TOLOI, 2004). Essa escolha é embasada, principalmente, nos gráficos das autocorrelações (FAC) e autocorrelações parciais (FACP). Ainda de acordo com o autor, o processo da identificação se dá em três partes:

1. Verificar, por com o auxílio de gráficos, se existe necessidade de uma transformação na série original, com o objetivo de estabilizar sua variância;
2. Tomar diferenças da série obtida, tantas vezes quanto necessárias para se obter uma série estacionária. Aqui se pode lançar mão de um teste para se verificar a existência de raízes unitárias no polinômio autorregressivo.
3. Identificar o processo resultante por meio das autocorrelações e autocorrelações parciais estimadas, que devem ser similares às autocorrelações e autocorrelações parciais teóricas.

2.4.1.1 Função de Autocorrelação e Função de Autocorrelação Parcial

Para Ehlers (2005), uma importante ferramenta para se identificar as propriedades de uma série temporal consiste de uma série de quantidades chamadas coeficientes de autocorrelação.

A autocorrelação e a autocorrelação parcial são medidas de associação entre valores de séries atuais e anteriores e indicam quais valores de série anteriores são mais úteis para prever os valores futuros.

A função de autocorrelação é uma medida da correlação entre as observações de uma série temporal que são separadas por k unidades de tempo (Z_t e Z_{t-1}).

A função de autocorrelação parcial é uma medida da correlação entre as observações de uma série temporal que são separadas por k unidades de tempo (Z_t e Z_{t-1}), após o ajuste para a presença de todos os outros termos de menor defasagem ($Z_{t-1}, Z_{t-2}, \dots, Z_{t-k-1}$).

A análise das funções de autocorrelação e autocorrelação parcial permite identificar se a série em análise atende ou não à condição de estacionariedade. Jacobs cita Matos (2000), o qual faz a seguinte observação a respeito da FAC:

Se os valores da função de autocorrelação iniciam-se próximos da unidade e declinam lenta e gradualmente à medida que aumenta a distância (número de defasagens k) entre os dois conjuntos de observações a que se referem, há indicações da existência de raiz unitária. A série de tempo é, portanto, não estacionária e segue um passeio aleatório. Se esses coeficientes declinam rapidamente à medida que tal distância se amplia, tem-se uma série com características de estacionariedade (MATOS, 2000, pág. 238 *apud* JACOBS, 2011).

As funções de autocorrelação e autocorrelação parcial devem ainda ser usadas em conjunto na fase da identificação de modelos. Através dos seus gráficos, denominados correlogramas, é possível identificar picos em cada *lag* (defasagem) para determinar se eles são significativos. Um pico significativo se estenderá para além dos limites significativos, o que indica que a correlação para esse *lag* não é igual a zero.

Quadro 1. Resumo das características teóricas da FAC e FACP para os modelos $AR(p)$, $MA(q)$ e $ARMA(p, q)$.

Modelo Teórico	Padrão Típico da FAC	Padrão Típico da FACP
$AR(p)$	Declina exponencialmente ou com padrão de onda senoidal amortecida, ou ambos.	Picos significativos através das defasagens p . Truncada em $k = p$.
$MA(q)$	Picos significativos através das defasagens q . Truncada em $k = p$.	Declina exponencialmente.
$ARMA(p, q)$	Declínio exponencial. Declinante a partir de $k = p$.	Declínio exponencial.

Fonte: Gujarati (2000); Fava (2000) *apud* Jacobs (2011).

O quadro 1 resume o comportamento das funções de autocorrelação e autocorrelação parcial para os modelos teóricos $AR(p)$, $MA(q)$ e $ARMA(p, q)$, permitindo comparar o modelo empírico da série em estudo com os modelos teóricos, o que facilita o processo de identificação daquele que melhor descreve a série temporal.

2.4.2 Estimação

Obtido o modelo ARIMA e seus valores de p , d e q e/ou P , D , Q , no caso de uma modelo ARIMA sazonal, deve-se agora estimar os parâmetros do modelo. Existem diferentes métodos para a estimativa dos parâmetros do modelo, e todos produzem resultados semelhantes (GUTIÉRREZ, 2003). Os mais comumente usados são o método dos mínimos quadrados (MQO) e o método de máxima verossimilhança (MMV). A equação abaixo apresenta a equação genérica do método de mínimos quadrados:

$$S(\hat{\phi}_1, \dots, \hat{\phi}_p, \hat{\theta}_1, \dots, \hat{\theta}_q) = \sum_{t=1}^n \hat{\varepsilon}_t^2, \quad (2.75)$$

em que:

$$\hat{\varepsilon}_t = \theta^{-1}(B)\phi(B)w_t. \quad (2.76)$$

Para determinar os parâmetros através da máxima verossimilhança deve-se admitir que os ruídos brancos são normalmente distribuídos. A equação abaixo apresenta a função de máxima verossimilhança:

$$L(\xi|Y, y_0) = (2\pi\sigma^2)^{-\frac{T}{2}} \exp \left\{ \frac{-1}{2\sigma^2} \sum_{t=1}^T (y_t - \phi_0 - \phi_1 y_{t-1})^2 \right\}. \quad (2.77)$$

Geralmente os pacotes estatísticos utilizados para séries temporais trazem o algoritmo para a estimação dos parâmetros.

2.4.3 Verificação

Nesta etapa verifica-se se o modelo representa a série de maneira adequada ou não. Em caso negativo, deve reiniciar o ciclo a fim de identificar outro modelo. A verificação de do modelo se dá através da análise dos resíduos e, segundo Jacobs (2011) através da avaliação da ordem do modelo. Nesta última, se verifica se o modelo está ou não superespecificado ou subespecificado, tendo algum ou ambos os parâmetros p e q maiores ou menores do que o devido.

Segundo Morettin e Toloi (2004), uma técnica que pode ser utilizada, caso haja suspeita de que um modelo mais elaborado, é o superajustamento. Ou seja, um modelo com parâmetros extras é estimado, em seguida deve-se examinar se estes parâmetros são significativos e se sua inclusão diminui significativamente a variância residual. Porém, entre dois modelos que se ajustam bem à uma série deve-se preferir aquele que tem menor número de parâmetros, isto é, deve-se atender ao princípio da parcimônia.

2.4.3.1 Análise de Resíduos

Para Ehlers (2005), se o modelo tiver um bom ajuste espera-se que os resíduos se distribuam aleatoriamente em torno de zero com variância aproximadamente constante e sejam não correlacionados. Evidência de correlação serial nos resíduos é uma indicação de que uma ou mais características da série não foi adequadamente descrita pelo modelo. Ainda de acordo com o autor,

Duas maneiras óbvias de se verificar a adequação do modelo consistem em representar graficamente os resíduos e o seu correlograma. O gráfico temporal poderá revelar a presença de dados discrepantes, efeitos de autocorrelação ou padrões cíclicos enquanto o correlograma permite uma análise mais detalhada da estrutura de correlação, indicando possíveis termos faltantes no modelo, ou seja, assim como em outros modelos estatísticos, a ideia é que os resíduos poderão identificar características que não foram adequadamente modeladas. Por exemplo, autocorrelações residuais significativas nas defasagens 1 ou 2, ou em defasagens sazonais são uma indicação de que mais termos médias móveis devem ser incluídos no modelo. (...) O mesmo vale para as autocorrelações parciais dos resíduos no que diz respeito à inclusão de termos autorregressivos (sazonais e não sazonais) (EHLERS, 2005, pág. 34).

2.4.3.2 Testes de Adequação do Modelo

Segundo Morettin e Toloi (2004), podem ser utilizados testes de diagnóstico de um modelo ajustado a uma série, baseados, geralmente, nas autocorrelações estimadas dos resíduos.

2.4.3.2.1 Teste de Autocorrelação Residual

O teste de autocorrelação de resíduos é empregado para verificar a dependência ou independência dos dados da série temporal. Consiste em analisar a relação entre variáveis defasadas e variáveis posteriores com base em diferentes intervalos de defasagem, evidenciando se os coeficientes são significativamente diferentes de zero (TÍBULO, 2014).

Segundo Morettin e Toloi (2004), estimados ϕ e θ , as quantidades:

$$\hat{a}_t = \hat{\theta}^{-1}(B)\hat{\phi}(B)W_t, \quad (2.78)$$

são chamadas resíduos estimados. Se o modelo for adequado, os \hat{a}_t deverão estar próximos dos a_t e, por consequência, deverão ser aproximadamente não correlacionados (aleatórios), isto é, as estimativas das autocorrelações dos resíduos deverão ser aproximadamente iguais a zero, sempre sob a suposição de que o modelo ajustado é apropriado. As autocorrelações são calculadas por:

$$\hat{r}_k = \frac{\sum_{t=k+1}^n \hat{a}_t \hat{a}_{t-k}}{\sum_{t=1}^n \hat{a}_t^2}. \quad (2.79)$$

2.4.3.2.2 Teste de Box-Pierce

O teste de Box-Pierce testa se um grupo de autocorrelações é significativamente diferente de zero através das chamadas estatísticas Q , dada por:

$$Q = n \sum_{k=1}^m r_k^2. \quad (2.80)$$

Na prática o número de autocorrelações amostrais é tipicamente escolhido entre 15 e 30. Se o modelo ajustado for apropriado então Q terá distribuição aproximadamente qui-quadrado com $m - p - q$ graus de liberdade. Assim, valores grandes de Q fornecem indicação contra a hipótese de que as autocorrelações são todas nulas, em favor da hipótese de que ao menos uma delas é diferente de zero (EHLERS, 2005). Segundo o autor, o teste de Box-Pierce não tem bom desempenho em amostras pequenas ou moderadas no sentido de que a distribuição se afasta da qui-quadrado, sendo sugerida uma variação deste por Ljung e Box.

2.4.3.2.3 Teste de Ljung-Box

Se o modelo for apropriado, a estatística do Teste de Ljung-Box é dada por:

$$Q(m) = n(n+2) \sum_{k=1}^m \frac{\hat{r}_k^2}{(n-k)}, \quad (2.81)$$

com distribuição aproximadamente qui-quadrado com $m - p - q$ graus de liberdade. A hipótese H_0 de ruído branco para resíduos é rejeitada para valores grandes de $Q(m)$.

2.4.3.3 Testes de Normalidade dos Resíduos

Se o modelo estiver correto os resíduos deverão ser independentes e identicamente distribuídos com distribuição normal $N(0,1)$. Para verificar se estes estão atendendo ao pressuposto da normalidade podem ser utilizados testes como o de Shapiro-Wilk e o de Kolmogorov-Smirnov. As outras verificações usuais sobre os resíduos também devem ser feitas, como por exemplo, um histograma com curva normal superposta e o gráfico de probabilidades normais (QQ-plot).

2.4.3.3.1 Teste de Shapiro-Wilk

O teste de Shapiro-Wilk foi proposto em 1965 e baseia-se no cálculo da estatística W , dada por:

$$W = \frac{b^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (2.82)$$

em que x_i são os valores da amostra ordenados. A constante b é determinada por:

$$b = \begin{cases} \sum_{i=1}^{\frac{n}{2}} a_{n-i+1} (x_{n-i+1} - x_i), & \text{se } n \text{ é par;} \\ \sum_{i=1}^{\frac{n+1}{2}} a_{n-i+1} (x_{n-i+1} - x_i), & \text{se } n \text{ é ímpar.} \end{cases} \quad (2.83)$$

As hipóteses para o teste são:

H_0 : A amostra provém de uma população normal;

H_1 : A amostra não provém de uma população normal.

Se o $p - \text{valor} < \alpha$, rejeita-se a hipótese de normalidade.

2.4.3.3.2 Teste de Kolmogorov-Smirnov

O teste de aderência de Kolmogorov-Smirnov pode ser utilizado para avaliar as seguintes hipóteses:

H_0 : Os dados seguem uma distribuição específica (distribuição normal);

H_1 : Os dados não seguem uma distribuição específica (distribuição normal).

Este teste observa a diferença máxima absoluta entre a distribuição normal (distribuição assumida para os dados) e a distribuição empírica dos dados. Como critério essa diferença é comparada a um valor crítico para dado nível de significância.

De acordo com Souza (2015), considerando-se uma amostra aleatória simples Z_1, Z_2, \dots, Z_n de uma população com função de distribuição acumulado contínua F_z desconhecida. A estatística utilizada para este teste é a diferença máxima absoluta entre as funções $F(z)$ e $F_n(z)$:

$$D_n = \max |F(z) - F_n(z)|. \quad (2.84)$$

Esta função corresponde à distância máxima vertical entre os gráficos de $F(z)$ e $F_n(z)$ sobre a amplitude dos possíveis valores de z . Em D_n temos que:

1. $F(z)$ representa a função de distribuição acumulada assumida para os dados (no caso a distribuição normal)
2. $F_n(z)$ representa a função de distribuição empírica dos dados.

Neste caso, procura-se testar a hipótese $H_0: F_n = F$ contra a hipótese alternativa $H_1: F_n \neq F$. Como se trata de um caso de análise da normalidade dos dados, assume-se que F seja a função de distribuição normal dos dados.

2.4.3.3.3 Teste de Jarque-Bera

O teste de Jarque-Bera é um teste assintótico (FERREIRA, 2006). As hipóteses a serem testadas são:

H_0 : Os dados seguem uma distribuição normal;

H_1 : Os dados não seguem uma distribuição não-normal.

O procedimento de teste consiste em calcular os valores de assimetria e da curtose da amostra e utilizar a estatística de teste definida por:

$$JB = N \left(\frac{\hat{\alpha}_1}{6} + \frac{(\hat{\alpha}_2 - 3)^2}{24} \right), \quad (2.85)$$

em que $\hat{\alpha}_1$ e $\hat{\alpha}_2$ são, respectivamente, os coeficientes amostrais de assimetria e curtose e N o tamanho da amostra. Jarque e Bera (1987) mostraram que sob a hipótese nula, assintoticamente a estatística JB segue uma distribuição χ^2 com dois graus de liberdade. A hipótese nula é rejeitada se $JB > \chi^2_{\alpha,2}$, em que $\chi^2_{\alpha,2}$ é o quantil de nível $1 - \alpha$ da distribuição χ^2 com dois graus de liberdade, ou ainda se o p -valor for inferior ao nível de significância proposto para o teste.

2.4.3.4 Critérios de Seleção de Modelos

De acordo com Emiliano (2010), não existem modelos verdadeiros. Há apenas modelos aproximados da realidade, que causam perda de informações. Deste modo, é necessário fazer a seleção do “melhor” modelo, dentre aqueles que foram ajustados, para explicar o fenômeno em estudo.

2.4.3.4.1 Critério de Akaike

O Critério de Akaike (AIC) foi proposto por Akaike (1973). Fazendo uso dos estimadores de máxima verossimilhança para os parâmetros do modelo, em que $l(\hat{\xi})$ é a função de log-verossimilhança maximizada e k é o número de parâmetros do modelo, o critério de Akaike (AIC) é dado por:

$$AIC = -2l(\hat{\xi}) + 2k. \quad (2.86)$$

2.4.3.4.2 Critério de Akaike Corrigido

Bozdogan (1987) propôs a seguinte correção para o Critério de Akaike (*apud* EMILIANO, 2010):

$$AIC = -2l(\hat{\xi}) + 2k + 2\frac{k(k+1)}{n-k-1}, \quad (2.87)$$

em que $l(\hat{\xi})$ é a função de log-verossimilhança maximizada e k representa o número de parâmetros.

2.4.3.4.3 Critério de Informação Bayesiano

O Critério de Informação Bayesiano (BIC) foi proposto por Schwarz (1978) e é dado por:

$$BIC = -2l(\hat{\xi}) + k\log(n), \quad (2.88)$$

em que $l(\hat{\xi})$ é a função de log-verossimilhança maximizada, k representa o número de parâmetros e n é o número de observações da amostra.

2.4.3.5 Medidas de Acuracidade

Tratam-se de estatísticas que permitem avaliar o ajuste de uma previsão aos dados originais por meio do cálculo de médias das diferenças (erros) entre os dados originais e as previsões em cada período da série temporal.

Baseiam-se na seguinte equação:

$$E = \sum_{t=1}^n |e_t(h)|. \quad (2.89)$$

Estes erros, porém, podem retornar valores positivos ou negativos, o que pode zerar o somatório dos erros. A fim de resolver este problema, diferentes maneiras de se calcular o somatório dos erros foram desenvolvidas e podem ser utilizadas para comparar o desempenho de diversos modelos (FURTADO, 2006). O melhor modelo para representar a série será o que apresentar os menores valores para as medidas de acuracidade.

Dentre as muitas medidas disponíveis destacam-se o Erro Absoluto Médio (MAE), o Erro Quadrático Médio (MSE), a Raiz Quadrada do Erro Quadrático Médio (RMSE) e o Erro Percentual Médio Absoluto (MAPE). Todas se baseiam nos cálculos dos erros: as diferenças entre os valores da série e os valores preditos pelas equações de tendência para cada período t da série. O Quadro 2 apresenta as citadas medidas de acuracidade e suas respectivas fórmulas.

Quadro 2. Principais Medidas de Acuracidade.

Medida	Equação
Erro Absoluto Médio	$MAE = \sum_{t=1}^n \frac{ Z_t - \hat{Z}_t }{n} \quad (2.90)$
Erro Quadrático Médio	$MSE = \sum_{t=1}^n \frac{(Z_t - \hat{Z}_t)^2}{n} \quad (2.91)$
Raiz Quadrada do Erro Quadrático Médio	$RMSE = \sqrt{\sum_{t=1}^n \frac{(Z_t - \hat{Z}_t)^2}{n}} \quad (2.92)$
Erro Percentual Absoluto Médio	$MAPE = \frac{100}{n} \sum_{t=1}^n \frac{ Z_t - \hat{Z}_t }{Z_t} \quad (2.93)$

onde:

Z_t é o valor da série no instante t ;

\hat{Z}_t é o valor previsto para o instante t ;

n é o número de períodos de tempos.

2.4.4 Previsão

Uma das formas de utilização de um modelo ajustado é fazer previsões de valores futuros (EHLERS, 2005). Uma vez identificado o processo estocástico que gera a série temporal em questão e os parâmetros do modelo, procede-se à etapa da previsão de valores futuros.

O interesse gira em torno de prever um valor Z_{t+h} , $h \geq 1$, supondo-se que têm-se observações $\dots, Z_{t-2}, Z_{t-1}, Z_t$, até o instante t , o qual é chamado origem das previsões. A previsão de origem t e horizonte h será denotada por:

$$\hat{Z}_t(h). \quad (2.94)$$

Com base nestas observações, pretende-se prever o valor da série Z no instante $t + h$, ou seja, os valores futuros $Z_{t+1}, Z_{t+2}, Z_{t+3}, \dots$. A previsão pra esse instante é dada pela esperança condicional de Z_{t+h} , dado o conjunto dos valores passados:

$$\hat{Z}_t(h) = [Z_{t+h}], \quad (2.95)$$

$$\hat{Z}_t(h) = E[Z_{t+h} | Z_t, Z_{t-1}, \dots]. \quad (2.96)$$

O erro de previsão h passos à frente será dado por:

$$e_t(h) = Z_{t+h} - \hat{Z}_t(h), \quad (2.97)$$

em que Z_{t+h} é o valor real e $\hat{Z}_t(h)$ é o valor previsto com origem em t e h passos à frente.

2.4.4.1 Atualização das Previsões

Segundo Ehlers (2005), as previsões podem ser atualizadas conforme novas observações da série forem sendo obtidas. Supondo-se que o valor X_{n+1} foi observado, a previsão para $t = n + 2$ ficará condicionada em X_1, \dots, X_n, X_{n+1} podendo ser facilmente atualizada para a nova origem $n + 1$.

3 METODOLOGIA

A título de exemplificar as etapas da metodologia Box e Jenkins foram considerados os dados da série histórica mensal correspondente ao Índice de Preços ao Consumidor Amplo (IPCA), compreendida no período de janeiro de 2007 a agosto de 2017. Os dados foram obtidos via internet no endereço eletrônico <http://www.ipeadata.gov.br/Default.aspx>, referente à base de dados mantida pelo Instituto de Pesquisa Econômica Aplicada (IPEA) e totalizam 128 observações, as quais constituem uma série temporal contínua. Destas 128 observações, 08 foram reservadas para validação e verificação da capacidade preditiva do modelo por meio da comparação com as previsões do modelo ajustado, de modo que, apenas 120 observações foram utilizadas nas fases de identificação, estimação e verificação do modelo.

Para obtenção das análises estatísticas dos dados, fez-se uso do *software* R versão 3.4.1 (R DEVELOPMENT CORE TEAM, 2017) e da sua IDE (Integrated Development Environment) R STUDIO, ambos gratuitos e de código aberto.

A modelagem de dados se deu com base na metodologia Box e Jenkins, observando o ciclo iterativo por ela proposto, como já mencionado na subseção 2.2.2 e melhor explicado na seção 2.4, isto é, a análise respeitou as etapas da identificação, da estimação de parâmetros, da verificação e, finalmente, o modelo foi utilizado para a previsão de valores futuros.

Em suma, foram realizados os seguintes procedimentos:

- 1) Análise gráfica da série original a fim de descrever o seu comportamento, e detectar indícios de volatilidade, tendência e/ou sazonalidade, observações atípicas e valores discrepantes;
- 2) Análise da condição de heterocedasticidade da série por meio de gráficos e medidas de dispersão; se necessário lançar mão de uma transformação de dados com a intenção de estabilizar a variância;
- 3) Análise gráfica das funções de autocorrelação (FAC) e autocorrelação parcial (FACP) da série original para, corroborando com o passo (1), verificar as condições de estacionariedade e sazonalidade da série, pode-se ainda fazer o uso de gráficos de decomposição da série em suas componentes e gráficos sazonais a fim de detectar a presença de padrão sazonal nos dados;
- 4) Através de testes estatísticos de tendência, raiz unitária, periodicidade e raiz unitária sazonal testar as suposições levantadas nas etapas anteriores com o objetivo de confirmar ou descartar a presença de tendência e sazonalidade;
- 5) Empregar diferenciações na série quantas forem necessárias com o objetivo de torná-la estacionária e/ou remover a sazonalidade;
- 6) Mais uma vez realizar testes estatísticos de tendência, raiz unitária e periodicidade para concluir se as componentes de tendência e sazonalidade foram removidas e a série atingiu a condição de estacionariedade;
- 7) Análise gráfica das funções de autocorrelação (FAC) e autocorrelação parcial (FACP) da série diferenciada a fim de identificar os possíveis modelos;
- 8) Estimação dos coeficientes dos modelos (S)ARIMA propostos no passo anterior;
- 9) Aplicação de testes de adequação do modelo para confirmar a aleatoriedade dos resíduos, isto é, a ausência de autocorrelação e se estes atendem à condição de ruído branco, análise das funções de autocorrelação dos resíduos;
- 10) Aplicação de testes de normalidade sobre os resíduos, a fim de verificar se estes atendem ao pressuposto da normalidade, análise gráfica do histograma com curva normal sobreposta e do gráfico das probalilidades QQ-Plot;

- 11) Escolha do melhor modelo ajustado com base nos critérios de informação disponíveis (AIC, AICC e BIC);
- 12) Escolha do melhor modelo preditivo com base nas medidas de acuracidade (MAE, RMSE), as quais permitem avaliar a capacidade de previsão de um modelo;
- 13) Análise de gráficos comparando as previsões entre os modelos ajustados;
- 14) Definição do melhor modelo ajustado e seus parâmetros estimados;
- 15) Realização de previsões h passos à frente com os modelo (S)ARIMA ajustado;
- 16) Comparar os resultados dos ajustes e das previsões do modelo (S)ARIMA com os valores da série histórica reservados para este fim.

3.1 CONSIDERAÇÕES SOBRE O IPCA

O Índice de Preços ao Consumidor Amplo (IPCA) é um índice criado para medir a variação de preços do mercado para o consumidor final e é utilizado pelo Banco Central como medidor oficial da inflação no Brasil. O governo usa o IPCA como referência para verificar se a meta estabelecida para a inflação está sendo cumprida. O índice é produzido e divulgado mensalmente desde 1980 pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Segundo o site do IBGE, o período de coleta do IPCA vai do dia 1º ao dia 30 ou 31, dependendo do mês. A pesquisa é realizada em estabelecimentos comerciais, prestadores de serviços, domicílios (a fim de verificar valores de aluguel) e concessionárias de serviços públicos. Os preços obtidos são efetivamente cobrados ao consumidor para pagamento à vista.

O indicador reflete a variação do custo de vida das famílias no Brasil e tem como alvo famílias com rendimento mensal entre um e quarenta salários mínimos, residentes nas regiões metropolitanas de São Paulo, Rio de Janeiro, Belo Horizonte, Porto Alegre, Curitiba, Salvador, Recife, Fortaleza, Belém e Goiânia, além do Distrito Federal.

São considerados os bens e serviços consumidos que formam a cesta de consumo, que são classificados em nove categorias de produtos e serviços, a saber, alimentação e bebidas, artigos de residência, comunicação, educação, habitação, saúde e cuidados pessoais, despesas pessoais, transportes e vestuário, os quais são subdivididos em outros itens. Ao todo são consideradas as variações de preços de 465 subitens.

A literatura mostra diversos estudos disponíveis acerca da modelagem da série referente ao IPCA, de maneira que os modelos (S)ARIMA, apesar de mais simples, têm se

mostrado superiores em termos de capacidade preditiva em relação às outras abordagens. São os casos dos trabalhos de Alencar (2006), Lorthiois (2009) e Bonno (2014). Bonno (2014) cita ainda os estudos de Newbold e Granger (1974), os quais também comprovaram a superioridade do modelo ARIMA. Lorthiois, por sua vez cita o trabalho de Stockton e Glassamn (1987), os quais consideram que um simples modelo ARIMA de inflação pode se tornar uma projeção de respeitável desempenho relativamente aos modelos teóricos baseados em especificações. Desta forma, os modelos ARIMA, desenvolvidos por Box e Jenkins, continuam a ser uma alternativa viável e satisfatória quando se trata de performance preditiva (MUTA *et al*, 2014).

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

A Tabela 1 apresenta as principais medidas descritivas obtidas a partir da série mensal correspondente ao Índice de Preços ao Consumidor Amplo (IPCA) compreendida no período de janeiro de 2007 a dezembro de 2016, possibilitando uma visão geral do comportamento da série em estudo.

Tabela 1. Medidas descritivas da série mensal correspondente ao Índice de Preços ao Consumidor Amplo (IPCA).

Medida Descritiva	Valor
Média	0,5035
Mediana	0,475
Moda	0,28
Variância	0,07163134
Desvio Padrão	0,2676403
Coefficiente de Variação	53,15597
Máximo	1,32
Mínimo	0,00
Curtose	0,5689544
Assimetria	0,6119938

Observando os resultados obtidos pode-se dizer que a média da série de dados foi de cerca de 0,50 com uma variabilidade de 0,27 em relação a esta. Observa-se ainda uma heterogeneidade do índice de preços durante o período, uma vez que o coeficiente de variação (relação entre desvio-padrão e média) é de aproximadamente 53,15%. Das 120 observações consideradas, o maior índice observado foi de 1,32, referente ao mês de março de 2015. Por outro lado, nota-se que a menor taxa foi de 0,00, referente ao mês de junho de 2010. A taxa mediana de 0,475 separa a série de preços ao meio, deixando 50% dos índices abaixo deste

valor e, por consequência, 50% das taxas acima do mesmo. O parâmetro de assimetria indica o formato das caudas de uma distribuição. Números positivos indicam uma cauda mais ampla à direita, enquanto números negativos indicam uma cauda mais ampla à esquerda. Desta forma, a assimetria positiva (0,612) indica que a distribuição possui uma cauda mais ampla à direita. A curtose de uma distribuição normal é igual a 3. Se a curtose exceder 3, a distribuição é denominada leptocúrtica. Se a curtose for inferior a 3, a distribuição é mais achatada e denominada platicúrtica. Observa-se, portanto, que a série é platicúrtica ($0,56 < 3$), ou seja, mais achatada em relação à curva normal.

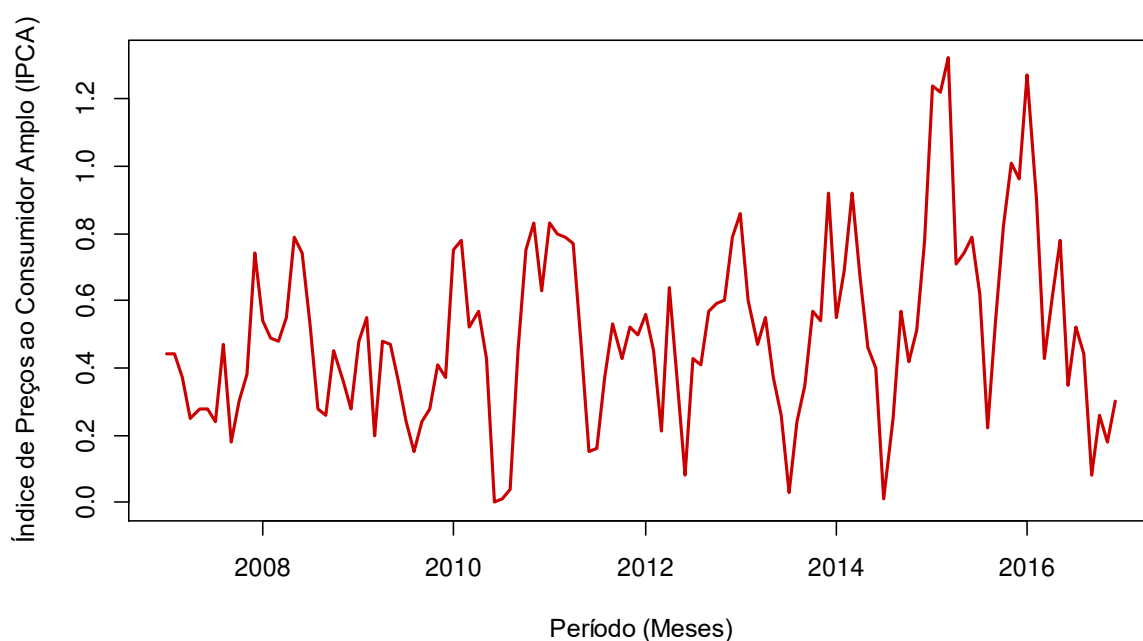


Figura 1. Representação gráfica da série original correspondente ao IPCA compreendida no período de janeiro de 2007 a dezembro de 2016.

Uma inspeção gráfica sugere a ausência de tendência, isto é, a série é, aparentemente, estacionária (Figura 1). Outra maneira de constatar estacionariedade da série temporal é visualizando o gráfico da função de autocorrelação (FAC) (Figura 2). A figura a mostra que as autocorrelações plotadas pela FAC decrescem de forma senoidal conforme descrito por Box e Jenkins. Esse é mais um indicativo de que a série temporal é estacionária.

Quanto à sazonalidade, observando o gráfico da série temporal e os gráficos das suas autocorrelações, identifica-se a presença de um forte padrão sazonal. No primeiro semestre, a inflação tende a cair, aumentando no segundo semestre. O gráfico da decomposição da série em suas principais componentes (Figura 3) também revela essa componente sazonal. Nota-se

ainda uma variância não constante no tempo. De maneira que pode ser necessária uma transformação não linear na série do Box-Cox a fim de estabilizá-la.

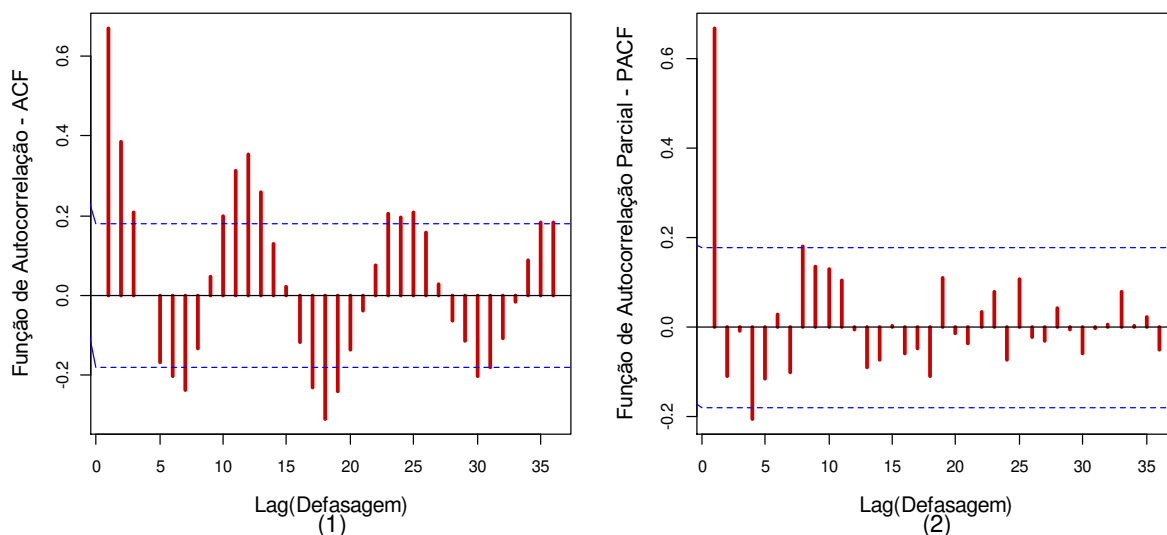


Figura 2. Representação gráfica das funções de (1) autocorrelação (FAC) e (2) autocorrelação parcial (FACP) da série original correspondente ao IPCA compreendida no período de janeiro de 2007 a dezembro de 2016.

Após a análise gráfica, o próximo passo é testar estatisticamente as percepções levantadas anteriormente, uma vez que, indícios de heterocedasticidade, não-estacionariedade e sazonalidade não são suficientes, estes apenas indicam um possível comportamento da série. Testes de raízes unitária não-sazonais e sazonais e testes de tendência e periodicidade são mais precisos e mais formais para a detecção de estacionariedade e sazonalidade.

Para testar a necessidade de uma transformação nos dados, o conselho de Ferreira (2011) é aplicar cada transformação para o maior e o menor dado de cada amostra. A amplitude dentro de cada amostra é determinada e a razão entre a maior e a menor amplitude é calculada. A transformação que produzir a menor razão é a selecionada.

Outra alternativa, também proposta por Morettin e Toloi (2004), é dividir os dados em subconjuntos, e, por meio de gráficos, nos quais se coloca no eixo das abscissas as médias e no eixo das ordenadas as variâncias de cada subconjunto para cada transformação, selecionar a transformação que apresentar a menor dispersão. Além disso, o coeficiente de variação é uma indicação razoável do efeito favorável das transformações de dados. Quando o valor do coeficiente de variação dos dados transformados for menor que o valor do coeficiente de variação dos dados originais, a transformação foi válida. Em caso contrário, não se justifica o seu uso.

Os procedimentos mencionados foram tomados para os dados em análise (Gráficos dados no Apêndice A). Foram testadas a transformação logarítmica e a transformação de raiz quadrada. Em um primeiro momento, a transformação de raiz quadrada se mostrou mais adequada, resolvendo o problema da heterogeneidade dos dados, uma vez que obteve a menor razão entre a amplitude máxima e a amplitude mínima, além de reduzir significativamente o coeficiente de variação da série (Tabelas 2 e 3). A transformação logarítmica, por outro lado não se mostrou adequada aos dados. Pelo contrário, aumentou ainda mais a sua variabilidade.

Tabela 2. Razão entre a amplitude máxima e a amplitude mínima da série correspondente ao IPCA para a série original, série de raiz quadrada e série de logaritmos.

	Série Original	Raiz Quadrada	Logarítmica
$Razão = \frac{Amp. máx}{Amp. mín}$	2,98	2,57	4,14

Tabela 3. Coeficientes de variação da série correspondente ao IPCA para série original, série de raiz quadrada e série de logaritmos.

	Série Original	Raiz Quadrada	Logarítmica
Coeficiente de Variação	53,16	30,71	90,60

No entanto, apesar de resolver o problema da variabilidade, ao estimar modelos para a série de dados para a qual a transformação de raiz quadrada foi aplicada, observou-se que esta produziu resíduos não normais para todos os modelos testados. Morettin e Toloí (2004), citando Nelson (1976), concluem que transformações não melhoram a qualidade da previsão. Citam ainda Makridakis e Hibon (1979), os quais verificaram que os dados transformados têm pouco efeito na melhoria da previsão e Granger e Newbold (1976), os quais observaram que a heterocedasticidade não afeta a adequação da previsão, pois não implica em estimadores viesados. Desta forma, diante dos resultados e do que Morettin e Toloí (2004), citando Nelson, Makridakis e Hibon; e Granger e Newbold, expõem, optou-se por trabalhar os dados originais, desprovidos de qualquer transformação.

Dado o forte padrão sazonal identificado graficamente, foram aplicados primeiramente testes para detectar a sazonalidade, uma vez que uma forte não estacionariedade sazonal pode fazer com que a série pareça não estacionária em seu comportamento não sazonal. A fim de verificar a presença dessa componente nos dados foram empregados os testes de Fisher, o qual apresentou p-valor inferior a 0,05, indicando a rejeição da hipótese nula de que não existe periodicidade na série ao nível de 5% de significância e o teste de Kruskal-Wallis, que, apresentando resultados semelhantes, rejeitou a hipótese nula de que não há sazonalidade,

com p -valor de 0,005649 e, conseqüente detecção de componente sazonal na série em estudo. Desse modo, confirmou-se a sazonalidade dos dados. Com o intuito de remover esta componente, a série recebeu uma diferença sazonal ($D = 1$), a fim de torná-la estacionária na parte sazonal (Figura 4). Em seguida, o teste de Fisher foi novamente aplicado, para confirmar a remoção da componente em estudo. Este apresentou p -valor superior a 0,05, indicando a não rejeição da hipótese nula de que não existe periodicidade, e a conseqüente remoção da sazonalidade através da aplicação de uma diferença sazonal.

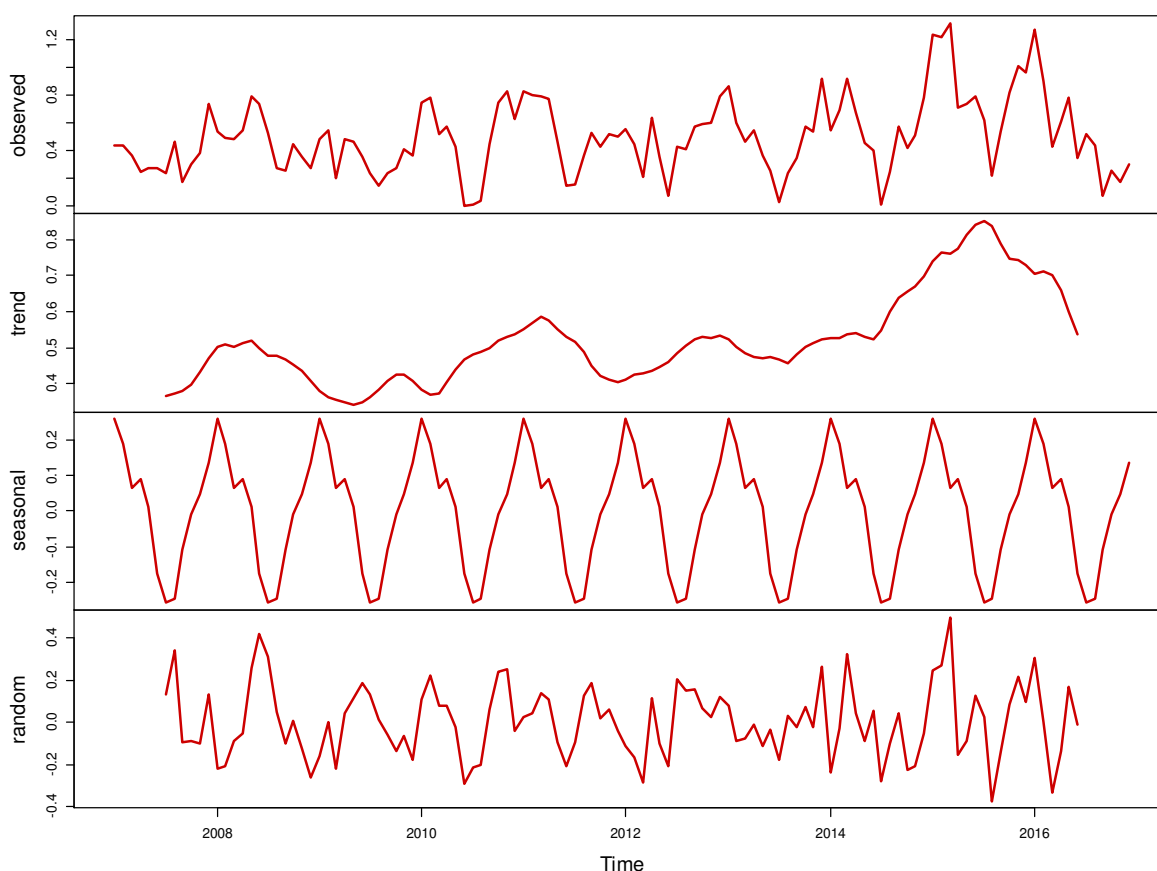


Figura 3. Representação gráfica da decomposição da série original correspondente ao IPCA compreendida no período de janeiro de 2007 a dezembro de 2016, nas componentes de tendência, sazonalidade e erro.

Em seguida foram aplicados testes para tendência e os testes de raiz unitária sobre a série diferenciada ($D = 1$). Para detectar a presença de tendência foi utilizado o teste de Cox-Stuart, o qual apresentou p -valor igual a 0,8919, superior a 0,05, indicando que não se pode rejeitar a hipótese nula de não existência de tendência com 95% de confiança. A fim de confirmar essa conclusão foram aplicados os testes de raiz unitária de Dickey-Fuller Aumentado (ADF), Phillips-Perron (PP) e Kwiatkowski-Phillips-Schmidt-Shin (KPSS).

Os resultados dos respectivos testes são apresentados na Tabela 4. Pelo teste de Dickey-Fuller Aumentado (ADF), a hipótese nula de raiz unitária não pode ser rejeitada ao nível de 0,05 de significância, isto é a série é não estacionária. O teste de Phillips-Perron (PP), rejeitou esta mesma hipótese com 95% de confiança e o teste Kwiatkowski-Phillips-Schmidt-Shin (KPSS), o qual possui por hipótese nula a ausência de raiz unitária, teve por resultado a não rejeição desta hipótese, atestando a estacionariedade da série em estudo.

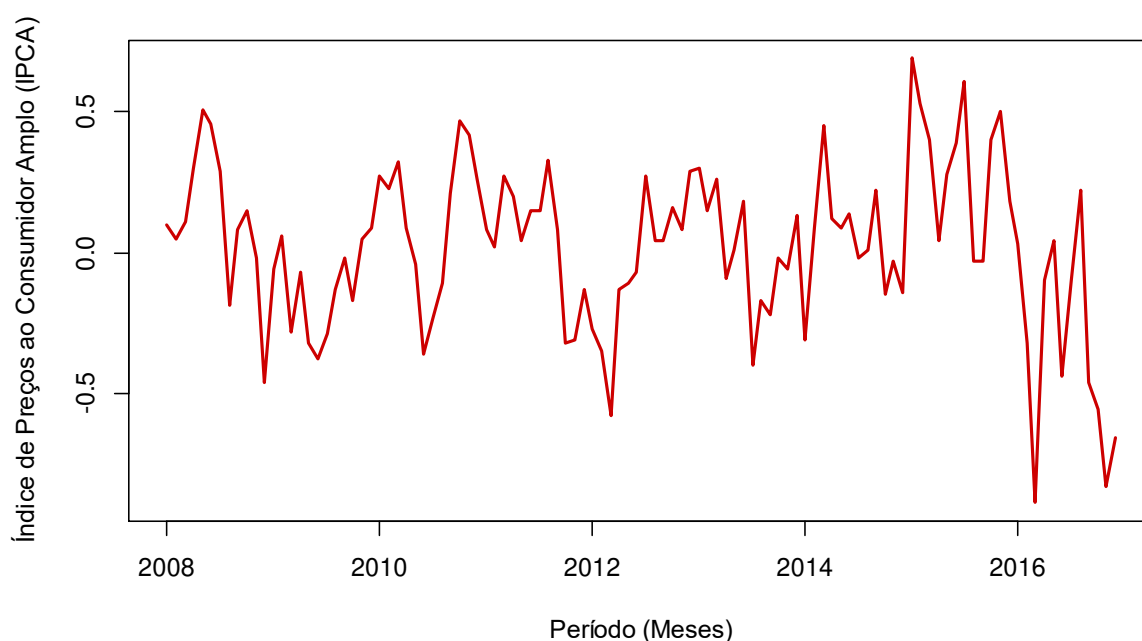


Figura 4. Representação gráfica da série correspondente ao IPCA compreendida no período de janeiro de 2007 a dezembro de 2016, submetida a uma diferenciação sazonal ($D = 1$).

Pode-se ainda, utilizando-se a função *ndiffs*, presente no pacote *forecast* do programa R, verificar a quantidade de diferenciações simples necessárias a fim de tornar a série estacionária na parte simples. A função *ndiffs* leva em conta o resultado dos testes de Dickey-Fuller Aumentado (ADF), Phillips-Perron (PP) e Kwiatkowski-Phillips-Schmidt-Shin (KPSS) e sugeriu a estacionariedade da série, indicando que a mesma não necessita de diferenciações simples.

Dessa forma, tomando como base o baixo poder estatístico do teste de Dickey-Fuller, os resultados dos testes de Cox-Stuart, Phillips-Perron (PP) e Kwiatkowski-Phillips-Schmidt-Shin (KPSS) e da função *ndiffs* e considerando que a diferenciação da série é uma transformação forte do processo que, se feita sem necessidade, induz a análises e modelagens equivocadas, a série foi considerada estacionária.

Tabela 4. Resultados dos Testes ADF, PP e KPSS aplicados à série mensal correspondente ao IPCA submetida a uma diferença sazonal.

Teste	ADF	PP	KPSS
P-valor	0,05628	0,01	0,1

Constatada a estacionariedade da série, procede-se a caracterização da mesma utilizando-se os devidos gráficos das autocorrelações (FAC) e autocorrelações parciais (FACP) de modo a identificar a ordem de termos autorregressivos (p) e médias móveis (q) que devem ser utilizados no modelo. Na Figura 5 têm-se os dois correlogramas.

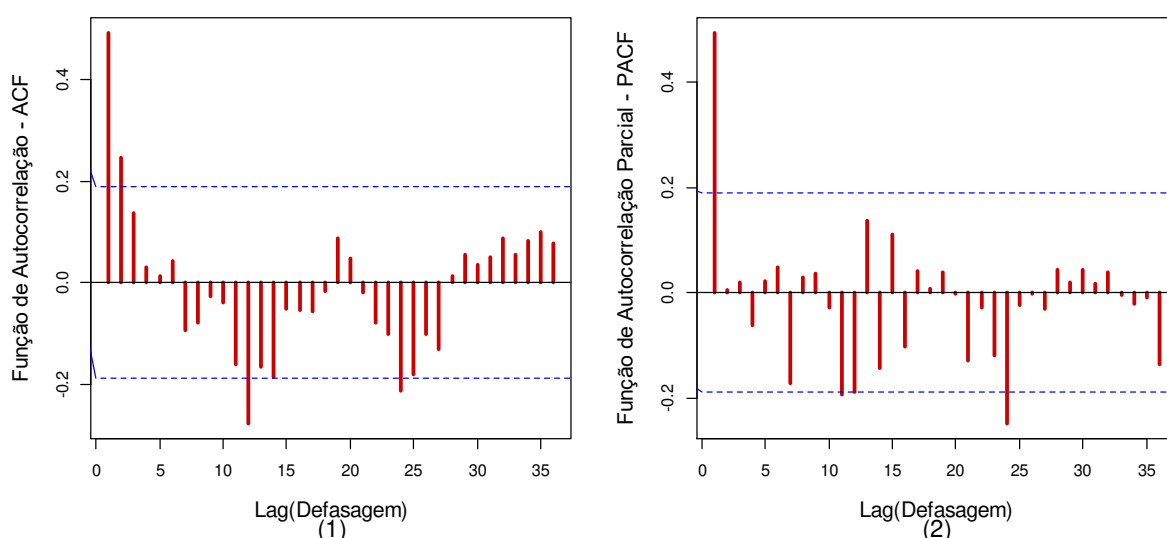


Figura 5. Representação gráfica das funções de (1) autocorrelação (FAC) e (2) autocorrelação parcial (FACP) da série mensal correspondente ao IPCA, compreendida no período de janeiro de 2007 a dezembro de 2016 submetida à uma diferença sazonal ($D = 1$).

Dado que as ordens dos operadores autorregressivos são identificáveis pela FACP e as ordens dos operadores de médias móveis pela FAC, através da análise dos correlogramas, nota-se um possível *lag* simples com correlação significativa na FAC, bem como um possível *lag* simples na FACP; na parte sazonal, têm-se um *lag* com correlação possivelmente significativa na FAC e um *lag* sazonal também possivelmente significativo na FACP. Desta forma, foram sugeridos inicialmente 08 modelos do tipo $SARIMA(p, d, q)(P, D, Q)_{12}$, os quais foram estimados e analisados quanto à significância dos parâmetros, a adequação do modelo e a normalidade dos resíduos. Apenas três modelos se mostraram aptos a representar a série dos Índices de Preços ao Consumidor Amplo (IPCA), a saber, os modelos:

- $SARIMA(0, 0, 1)(1, 1, 0)_{12}$;
- $SARIMA(1, 0, 0)(0, 1, 1)_{12}$;

- *SARIMA* (1, 0, 0) (1, 1, 0)₁₂.

A tabela 5 apresenta suas respectivas estatísticas. Verifica-se, por meio dos intervalos de confiança, que todos os parâmetros de cada modelo são estatisticamente diferentes de zero. Tal significância também foi comprovada por meio do teste *Z* para os coeficientes a 5% de significância.

Tabela 5. Estatísticas dos modelos *SARIMA* (*p, d, q*) (*P, D, Q*)₁₂ estimados para a série mensal correspondente ao IPCA.

SARIMA (0, 0, 1) (1, 1, 0)					
Parâmetros	Estimativas	Erro Padrão	Intervalo de Confiança		$\hat{\sigma}^2$
			2,5%	97,5%	
MA (1)	0,5001	0,0715	0,3600678	0,6402004	0,05473
SAR (1)	-0,3778	0,1023	-0,5783834	-0,1771961	
SARIMA (1, 0, 0) (0, 1, 1)					
Parâmetros	Estimativas	Erro Padrão	Intervalo de Confiança		$\hat{\sigma}^2$
			2,5%	97,5%	
AR (1)	0,5438	0,0826	0,4541761	0,7580581	0,0336
SMA (1)	-1,0000	0,1426	-1,479794	-0,520242	
SARIMA (1, 0, 0) (1, 1, 0)					
Parâmetros	Estimativas	Erro Padrão	Intervalo de Confiança		$\hat{\sigma}^2$
			2,5%	97,5%	
AR (1)	0,5428	0,0832	0,3796720	0,7059114	0,05269
SAR (1)	-0,3484	0,1064	-0,5569939	-0,1397218	

Após a estimação e validação dos parâmetros, os resíduos foram analisados quanto às características de independência, através dos testes de Ljung-Box e Box-Pierce; e normalidade, por meio dos testes de Shapiro-Wilk, Kolmogorov-Smirnov e Jarque-Bera (análise gráfica disponível no Apêndice C). Os resultados são dados na Tabela 6 e indicam a não rejeição da hipótese nula de independência dos resíduos para os testes de Ljung Box e Box Pierce ao nível de 5% de significância, isto é, os resíduos são não correlacionados ao longo do tempo.

Tabela 6. Resultado dos *p-valores* para os testes de Ljung-Box, Box-Pierce, Shapiro-Wilk, Kolmogorov-Smirnov e Jarque-Bera para os modelos ajustados à série mensal correspondente ao IPCA.

Modelo	Teste dos Resíduos: <i>P-valores</i>				
	Ljung-Box	Box-Pierce	Shapiro-Wilk	Kolmogorov-Smirnoff	Jarque-Bera
SARIMA (0, 0, 1) (1, 1, 0)	0,0734	0,1709	0,4607	0,1085	0,5462
SARIMA (1, 0, 0) (0, 1, 1)	0,6912	0,8154	0,8792	0,5332	0,9114
SARIMA (1, 0, 0) (1, 1, 0)	0,2074	0,3863	0,1208	0,1545	0,1694

Analogamente, os testes de normalidade, os quais têm por hipótese nula a distribuição normal da série, apresentaram *p-valores* superiores ao nível de significância fixado, 0,05, de modo que não existem evidências para se rejeitar esta hipótese. Os resíduos podem, portanto, ser considerados normalmente distribuídos.

Tabela 7. Resultados dos Critérios de Informação AIC, AICC e BIC e das Medidas de Acurácia MAE e RMSE para os modelos ajustados à série mensal correspondente ao IPCA.

Modelo	AIC	AICC	BIC	MAE	RMSE
SARIMA (0, 0, 1) (1, 1, 0)	-1,864794	-1,845229	-2,795107	0,1695128	0,2219419
SARIMA (1, 0, 0) (0, 1, 1)	-2,343127	-2,323562	-3,273440	0,1372884	0,1769676
SARIMA (1, 0, 0) (1, 1, 0)	-1,896570	-1,877000	-2,826880	0,1647563	0,2177547

Diante dos resultados, todos os modelos foram bem ajustados. Como critério de comparação e seleção do melhor modelo foram utilizados os Critérios de Informação de Akaike (AIC), Akaike Corrigido (AICC) e Bayesiano (BIC), além da Raiz Quadrada do Erro Quadrático Médio (RMSE), o Erro Absoluto Médio (MAE) e a variância (Tabela 7). Segundo os critérios AIC, AICC e o BIC, o melhor modelo é o *SARIMA (1, 0, 0) (1, 1, 0)₁₂*. Porém, de acordo com a variância, o RMSE e o MAE, que medem o desempenho da previsão, o modelo mais adequado é o *SARIMA (1, 0, 0) (0, 1, 1)₁₂*. Analisando o gráfico das previsões encontradas pelos três modelos (Figura 6), aparentemente, os modelos *SARIMA (0, 0, 1) (1, 1, 0)₁₂* e *SARIMA (1, 0, 0) (1, 1, 0)₁₂*, com destaque para o modelo *SARIMA (1, 0, 0) (1, 1, 0)₁₂*, acompanharam melhor o comportamento da série original do IPCA, isto é, apresentaram melhor ajuste a esta em comparação ao modelo *SARIMA (1, 0, 0) (0, 1, 1)₁₂*. Porém, este foi o que apresentou as melhores previsões.

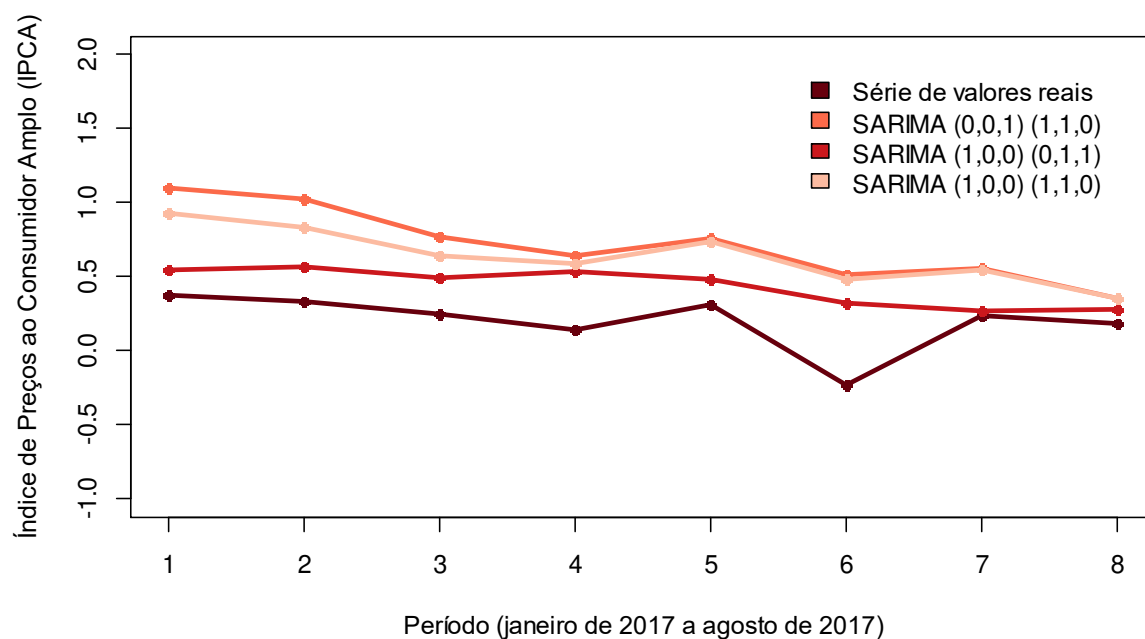


Figura 6. Comparação entre os valores reais da série mensal IPCA e os valores estimados pelos modelos SARIMA no período de janeiro a agosto de 2017.

As previsões fornecidas pelo modelo $SARIMA(1, 0, 0)(0, 1, 1)_{12}$ 8 passos à frente são dadas na Tabela 8.

Tabela 8. Previsões $h = 8$ para o modelo $SARIMA(1, 0, 0)(0, 1, 1)_{12}$.

Período	Índice Previsto	Índice Observado	Limite Inferior 5%	Limite Superior 95%
Jan / 2017	0,5457876	0,3800000	0,161622840	0,9299524
Fev / 2017	0,5669356	0,3300000	0,118316641	1,0155546
Mar / 2017	0,4950718	0,2500000	0,024988345	0,9651552
Abr / 2017	0,5337732	0,1400000	0,056046518	1,0114999
Mai / 2017	0,4866424	0,3100000	0,006138632	0,9671462
Jun / 2017	0,3232532	-0,2300000	-0,158265827	0,8047722
Jul / 2017	0,2673215	0,2400000	-0,214567295	0,7492102
Ago / 2017	0,2784005	0,1900000	-0,203616722	0,7604178

Apesar de todos os modelos mostrarem-se bem ajustados à série mensal dos Índices de Preço ao Consumidor Amplo (IPCA), como o objetivo principal da análise de séries temporais é a previsão, foi selecionado o modelo que apresentou menor variância e os menores valores para as medidas do Erro Absoluto Médio (MAE) e a Raiz Quadrada do Erro Quadrático Médio (RMSE). Vale pontuar que, embora o uso de critérios de informação seja útil para identificar o modelo que melhor equilibra qualidades de ajuste de parcimônia, os mesmos não são de caráter eliminatório. Nesse sentido, através das medidas MAE e RMSE,

selecionou-se, dentre os modelos ajustados, aquele que fornece as melhores previsões para fora da amostra, a saber, o modelo $SARIMA(1, 0, 0)(0, 1, 1)_{12}$ cuja equação é dada abaixo:

$$IPCA_t = 0,0022 + 0,5438IPCA_{t-1} + \varepsilon_t + \varepsilon_{t-12}, \quad (4.1)$$

em que 0,0022 é uma constante significativa no modelo.

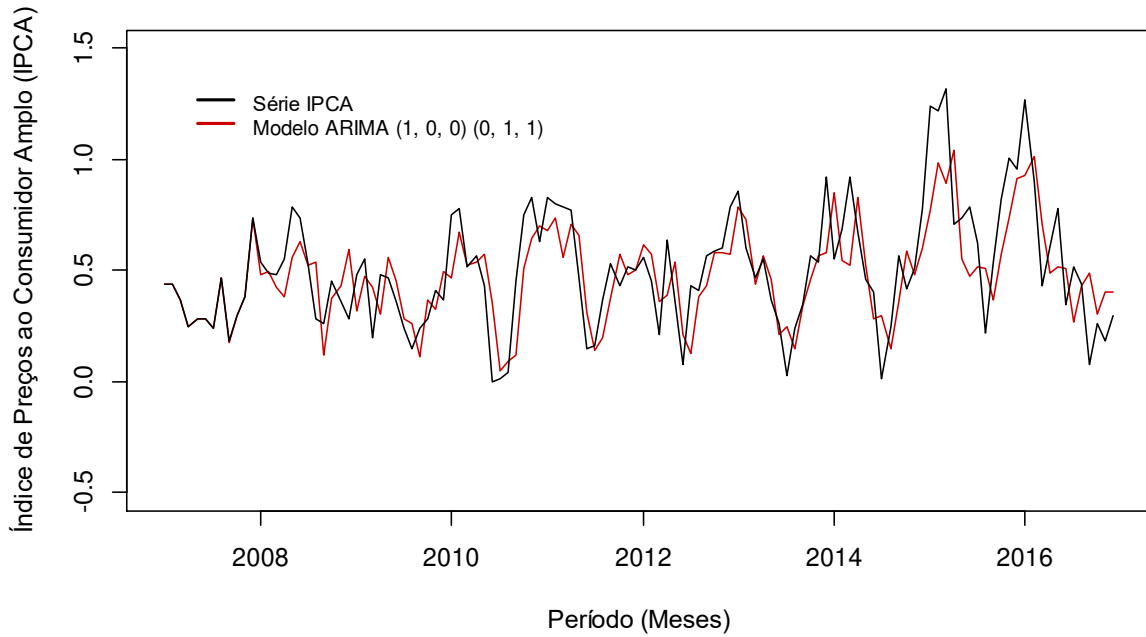


Figura 7. Ajuste do Modelo $SARIMA(1, 0, 0)(0, 1, 1)_{12}$ à série IPCA.

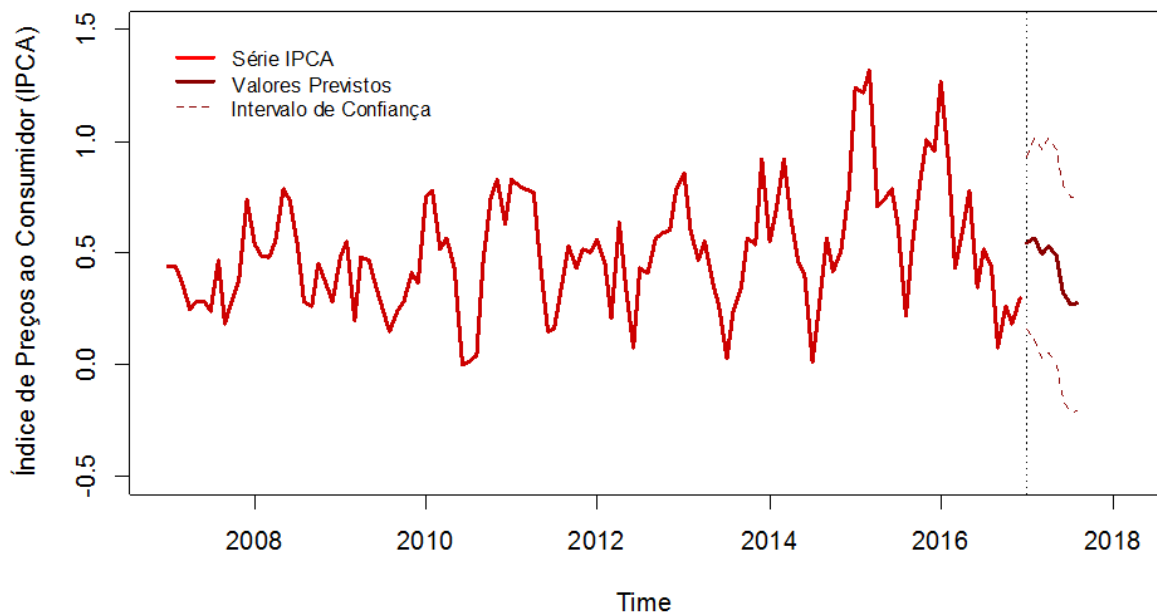


Figura 8. Representação gráfica da série mensal correspondente ao IPCA compreendida no período de janeiro de 2007 a dezembro de 2016, a série de índices previstos pelo modelo $SARIMA(1, 0, 0)(0, 1, 1)_{12}$ para $h = 8$ e o intervalo de confiança para as previsões

O gráfico da modelo $SARIMA(1, 0, 0)(0, 1, 1)_{12}$ ajustado à série correspondente ao Índice de Preços ao Consumidor Amplo (IPCA) é dado na Figura 7, ao passo que a Figura 8 refere-se ao gráfico da série mensal correspondente aos Índices de Preços ao Consumidor Amplo (IPCA) compreendida no período de janeiro de 2007 a dezembro de 2016 juntamente com a série de índices previstos pelo modelo $SARIMA(1, 0, 0)(0, 1, 1)_{12}$ para $h = 8$ e o intervalo de confiança para as previsões obtidas.

5 CONCLUSÃO

O objetivo deste trabalho foi reunir aspectos principais sobre a metodologia de alguns modelos lineares utilizados para previsão em séries temporais, enfocando os modelos ARIMA e a Metodologia Box e Jenkins, a qual, segundo Souza, L. V., (2006) e Figueredo (2008), é a mais importante, eficiente e utilizada metodologia para a previsão em séries temporais.

Objetivando, ainda, a demonstração de uma aplicação prática dos modelos Box e Jenkins, a referida metodologia foi aplicada à série de dados correspondente ao Índice de Preços ao Consumidor Amplo (IPCA), considerado o índice oficial da inflação no Brasil, obedecendo-se o ciclo iterativo da identificação, estimação, verificação e previsão.

Foram propostos inicialmente 08 modelos, no entanto apenas 03, foram aprovados na etapa da verificação, se mostrando aptos a representar a série de dados em questão. Dentre os três modelos aptos, em termos de capacidade preditiva, o modelo $SARIMA(1, 0, 0)(0, 1, 1)_{12}$ obteve o melhor desempenho.

Quanto à dinâmica da inflação medida pelo IPCA, os três modelos conseguiram captá-la de forma razoável, com destaque para o modelo $SARIMA(1, 0, 0)(1, 1, 0)_{12}$. Por se tratar de modelos univariados, obviamente, os modelos SARIMA excluem efeitos de outras variáveis ou de eventuais choques que podem ocorrer sobre a inflação mensal, não sendo capazes, portanto, de prever mudanças de comportamento bruscas de outras naturezas que tenham capacidade de explicação na série. Nesse contexto, para tornar a previsão mais acurada é necessário incorporar à modelagem outras variáveis intrínsecas ao processo.

Tratando-se de um estudo inicial, o modelo sugerido foi escolhido a fim de ilustrar a aplicação da metodologia Box e Jenkins, porém apresentou pouca aderência ao comportamento da série, sendo necessária a aplicação de uma metodologia que contemple à modelagem variáveis exógenas como o hiato do produto, as expectativas da inflação e os movimentos da taxa de câmbio, as quais, segundo a teoria econômica, influenciam no índice

da inflação. Desta forma, a fim de que essas variáveis possam ser incorporadas, sugere-se, para análises futuras, estudos aprofundados sobre os modelos SARIMAX.

REFERÊNCIAS

- ALENCAR, B. K. **Modelos de previsão da inflação: uma análise comparativa no curto prazo**. Dissertação (Mestrado em Economia). Instituto Brasileiro de Mercado de Capitais – IBMEC/RJ, 2006.
- BARBOSA, C. M. S. **Gestão da cadeia de abastecimento de bens de consumo baseada em modelos de previsão lineares**. Dissertação (Mestrado em Logística) – Instituto Politécnico do Porto – APNOR/PORTUGAL, 2014.
- BONNO, S. J. P. **Previsão de Inflação Utilizando Modelos de Séries Temporais**. Dissertação (Mestrado em Economia) – Fundação Getúlio Vargas – FGV, 2014.
- BRAGA, J. De M. **Ajustamento nos Mercados de Fatores, Raiz Unitária e Histerese na Economia Americana**. Texto para Discussão, Faculdade de Economia, Universidade Federal Fluminense – UFF/RJ, 2008.
- CANOVA, F.; HANSEN, B. E. **Are Seasonal Patterns Constant Over Time? A Test for Seasonal Stability**, *Journal of Business & Economic Statistics* 13, 237-252.
- CARVALHO, P. L. C.; FERRAZ, M. I. F.; SÁFADI T. **Verificação da existência da componente sazonal na série do custo da cesta básica do município de Lavras, MG, utilizando o método X-12 ARIMA**. *Organizações Rurais & Agroindustriais*, Lavras, v. 8, 268-278, 2006.
- CORDEIRO, B. A. F.; RANIERI, J. R.; RODRIGUES, N. A. C.; FERREIRA, V. T. de S. **Aplicação do Modelo de Média Exponencial para a Realização da Previsão de Demanda do Setor Leiteiro no Brasil**. *XXXV Encontro Nacional De Engenharia de Produção: Perspectivas Globais para a Engenharia de Produção*. Fortaleza/CE, 2015.
- CORTEZ, P. A. R., **Modelos Inspirados na Natureza para a Previsão de Séries Temporais**. 188 f. Tese (Doutorado em Informática) – Departamento de Informática, Universidade do Minho – Braga, 2002.
- CRUZ, R. C. **Uso de Funções de Transferência e da Transformada Wavelet aplicadas a Sinais Meteorológicos da Floresta de Caxiuanã**. Dissertação (Mestrado em Estatística) – Universidade Federal do Pará – UFPA/PA, 2008.
- EHLERS, R. S. **Análise de Séries Temporais**. Departamento de Matemática Aplicada e Estatística –UFPR (2005). Disponível em <<http://leg.est.ufpr.br/~ehlers/notas>>. Acesso em 05/02/2017.
- EMILIANO, P. C., VEIGA, E. P.; VIVANCO, M. J. F.; MENEZES, F. S. De. **Crítérios de informação de Akaike versus Bayesiano: Análise Comparativa**. *19º Simpósio Nacional de Probabilidade e Estatística* (2010).
- FERREIRA, M. R. P. **Análise da Sensibilidade dos Testes de Normalidade de Jarque-Bera e Lilliefors em Modelos de Regressão Linear**. *Revista Matemática Estatística*, v. 24, n. 4, 89-98, 2006.

FERREIRA, P. V. **Introdução à Estatística Experimental**. Centro de Ciências Agrárias (CECA) – Universidade Federal de Alagoas – UFAL, 2011.

FIGUEREDO, C. J. **Previsão de Séries Temporais Utilizando a Metodologia Box & Jenkins e Redes Neurais para Inicialização de Planejamento e Controle de Produção**. Dissertação (Mestrado em Ciências) – Universidade Federal do Paraná – UFPR, 2008.

FURTADO, M. R. **Aplicação de um modelo de previsão da demanda total nos credenciados Belgo Pronto**. Monografia (Graduação em Engenharia de Produção) – Universidade Federal de Juiz de Fora – UFJF, 2006.

GARBER, R.; PERES N.; PR; VALENTIN, J. L.; FERNANDEZ, F. **Análise de séries temporais**. *Ecologia Brasiliensis*: tópicos em tratamento de dados biológicos. Rio de Janeiro: UERJ 2 (1995): 91-118.

GUJARATI, D. **Econometria básica**. Rio de Janeiro: Elsevier, 2006. 812p.

GUTIÉRREZ, J. L. C. **Monitoramento da instrumentação da barragem Corumbá I por redes neurais e modelos de Box & Jenkins**. Dissertação (Mestrado em Engenharia Civil) – Pontifícia Universidade Católica do Rio de Janeiro – PUC/RJ, 2003.

HENNING, E.; MARTIN, A. C.; WALTER, O. M. F. C.; HACKENBERG, A. M.; LOUREIRO, K. da C. **Previsão de vendas de motocicletas aplicando modelos SARIMA, suavização exponencial e combinação de previsões**. *Revista de Engenharia e Tecnologia* 7.3 (2015): Páginas-123.

IBGE. Instituto Brasileiro de Geografia e Estatística. **Índice Nacional de Preços ao Consumidor Amplo – IPCA e Índice Nacional de Preços ao Consumidor – INPC**. Disponível em: <<http://www.ibge.gov.br>>. Acesso em: agosto de 2017.

IPEA. Instituto de Pesquisa Econômica Aplicada. **Ipeadata – Inflação: IPCA**. Disponível em: <<http://www.ipeadata.gov.br>>. Acesso em: agosto de 2017.

JACOBS, W. **Modelos de suavização exponencial, ARIMA e redes neurais artificiais: um estudo comparativo para a previsão de demanda de produtos**. Monografia (Bacharelado em Engenharia de Produção) – Centro Universitário – UNIVATES/RS, 2011.

KLIDZIO, R. **Modelos de previsão aplicados ao controle de qualidade com dados autocorrelacionados**. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal de Santa Maria – UFSM, 2009.

LORTHIOIS, A. **Modelo de Previsão de Inflação no Brasil**. Dissertação (Mestrado em Economia e Econometria) – Fundação Getúlio Vargas – FGV, 2009.

MARGARIDO, M. A.; JUNIOR, H. de M. **Teste para mais de uma raiz unitária: uso do software SAS® na elaboração de uma rotina para o teste Dickey-Pantula**. *Pesquisa & Debate. Revista do Programa de Estudos Pós-Graduados em Economia Política*. ISSN 1806-9029 17.1 (29) (2006).

MARTIN, A. C.; HENNING, E.; WALTER, O. M. F. C.; KONRATH, A. C. **Análise de séries temporais para previsão da evolução do número de automóveis no Município de Joinville.** *Revista ESPACIOS*, Vol. 37 (Nº 06) Año 2016 (2016).

MARTÍNEZ, R.. O.; ZAMPROGNO, B. **Comparação de Algumas Técnicas de Previsão em Análise de Séries Temporais.** *Revista Colombiana de Estadística*, v. 26, n. 2, 129-157, 2003.

MIRANDA, I. P. De. **Comparação de diferentes Métodos de Previsão em Séries Temporais com valores discrepantes.** Monografia (Bacharelado em Estatística) – Universidade Federal de Juiz de Fora – UFJF/MG, 2014.

MOREIRA, R. J. S. **Análise das Características de Ruído em Séries Temporais GPS.** Dissertação (Mestrado em Engenharia Geográfica) – Universidade de Lisboa, 2013.

MORETTIN, P. A.; TOLOI, C. M. C. **Análise de Séries Temporais.** São Paulo: Edgard Blucher LTDA, 2004, 535 p.

MUTA, A. A.; FIGUEIREDO, C.C. **Um modelo de previsão para o IPCA: como a modelagem econométrica pode ser útil para estimá-lo antes das apurações oficiais.** *Seminário de Iniciação Científica, ESPM/SP*. São Paulo/SP, 2014.

R DEVELOPMENT CORE TEAM (2017). **R: a language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <<http://www.r-project.org>>. Acesso em: junho de 2017.

REIS, M. M. **Análise de Séries Temporais.** Disponível em: <www.inf.ufsc.br/~marcelo/Cap4.pdf> Acesso em: agosto de 2017.

RELVAS, T. R. S. **Previsões com séries temporais: cálculo de coeficientes sazonais com regressão linear.** *Anais do Congresso Brasileiro de Custos-ABC*. 1998.

RODRIGUÉZ, S.; MASSA, F. **¿Estacionalidad Determinística em los Componentes Del Producto?** Instituto de Estadística, 2012.

SALVIANO, M. F.; GROppo, J. D.; PELLEGRINO, G. Q. **Análise de Tendências em Dados de Precipitação e Temperatura no Brasil.** *Revista Brasileira de Meteorologia*, v. 31, n. 1, 64-73, 2016.

SOUZA, F. M. **Modelos Box & Jenkins aplicados a previsão de demanda de leitos hospitalares.** Monografia (Especialização em Estatística e Modelagem Quantitativa) – Universidade Federal de Santa Maria – UFSM/RS, 2006.

SOUZA, L. V. **Programação Genética e Combinação de Preditores para Previsão de Séries Temporais.** Tese (Doutorado em Ciências) – Universidade Federal do Paraná – UFPR, 2006.

SOUZA, R. B. de. **Análise de séries temporais e previsão: um estudo de aplicação prática da metodologia Box & Jenkins no consumo mensal de energia elétrica na região norte do**

Brasil. Monografia (Bacharelado em Estatística) – Universidade Federal de Rondônia – UNIR/RO, 2015

TIBULO, C. Modelos de séries temporais aplicados a dados de umidade relativa do ar. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal de Santa Maria – UFSM/RS, 2014.

ZAMPROGNO, B. Estimação e testes de processos estacionários e não estacionários sazonais com longa dependência. Dissertação (Mestrado em Estatística) – Universidade Federal de Pernambuco – UFPE/PE, 2004.

APÊNDICE A

Para verificar a necessidade de transformação nos dados a fim de resolver o problema da variância não constante, o conselho de Morettin e Toloi (2004) e Ferreira (2011) é dividir os dados em subconjuntos, e, por meio de gráficos, nos quais se coloca no eixo das abscissas as médias e no eixo das ordenadas as variâncias de cada subconjunto para cada transformação, selecionar a transformação que apresentar a menor dispersão. Os gráficos mencionados para a série referente ao Índice de Preços ao Consumidor Amplo (IPCA) são dados abaixo:

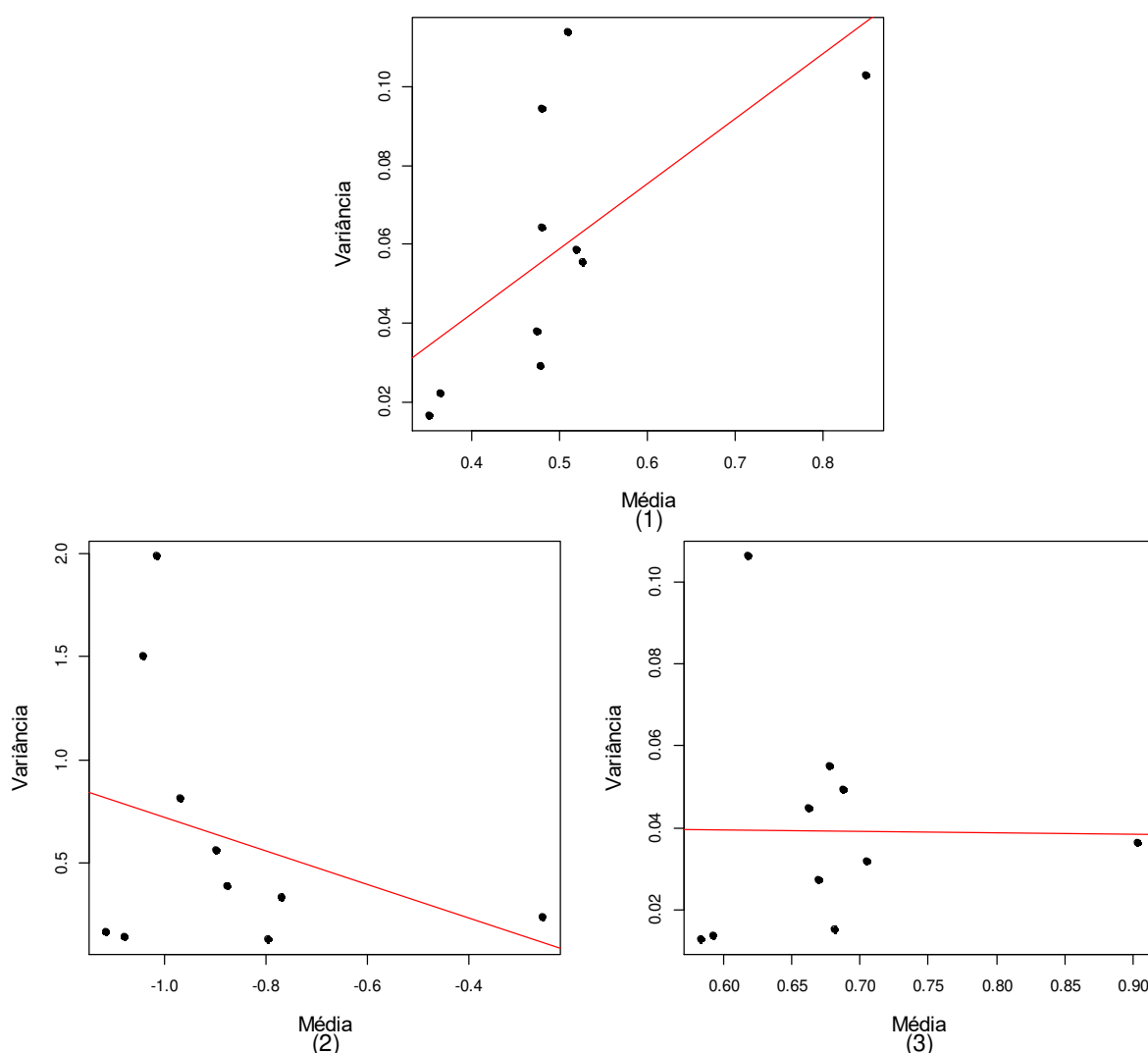


Figura 9. Representação gráfica da média versus variância da série correspondente ao IPCA compreendida no período de janeiro de 2007 a dezembro de 2016: (1) Série Original, (2) Transformação Logarítmica e (3) Transformação Raiz Quadrada.

APÊNDICE B

Gráficos Auxiliares na Identificação da Sazonalidade nos Dados

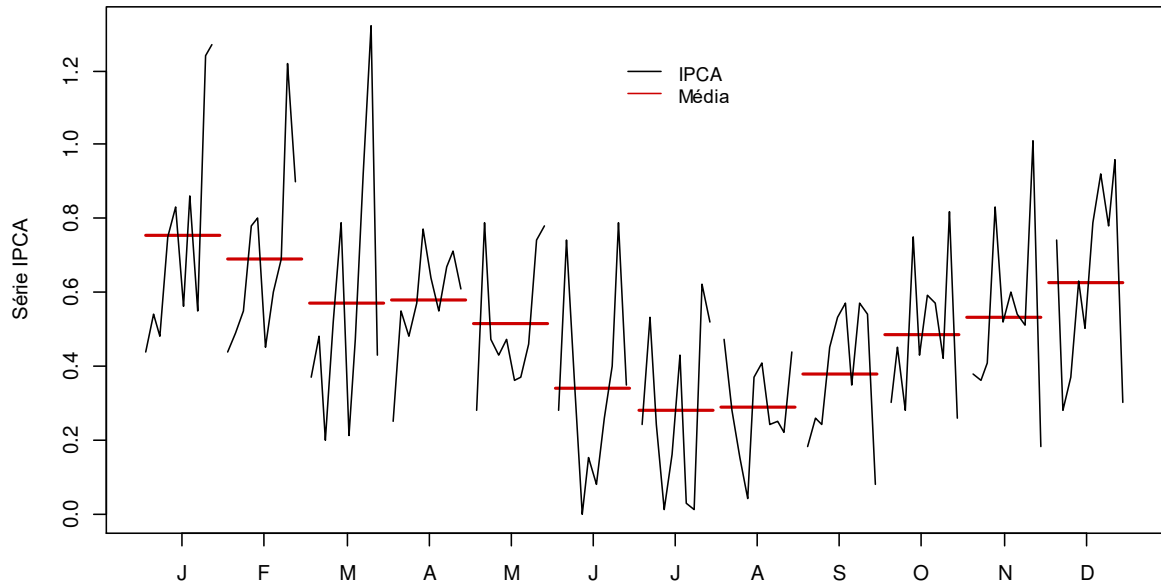


Figura 10. Representação Gráfica do IPCA por mês e do IPCA médio por mês.

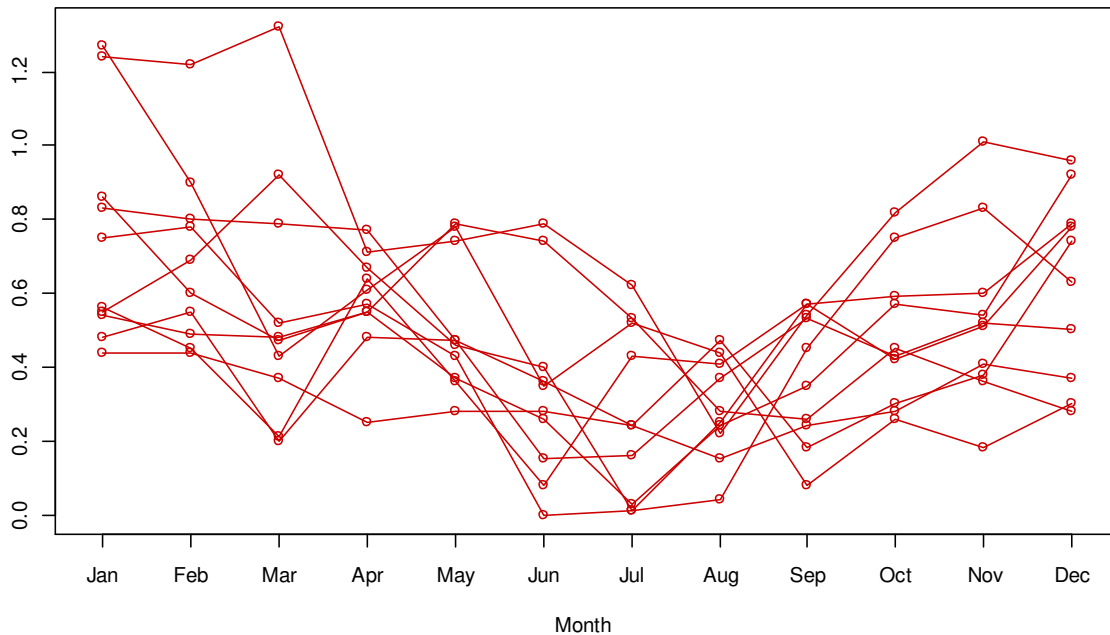


Figura 11. Gráfico Sazonal.

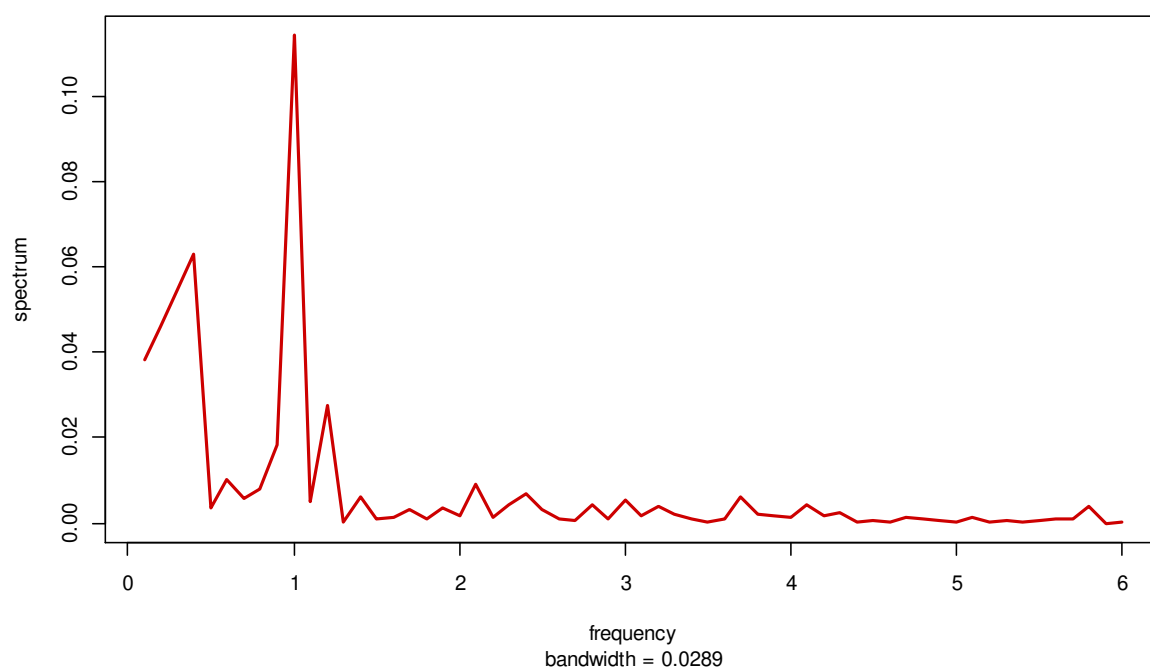


Figura 12. Periodograma da série referente ao IPCA.

APÊNDICE C

Análise Gráfica dos Resíduos dos Modelos $SARIMA(0,0,1)(1,1,0)_{12}$, $SARIMA(1,0,0)(0,1,1)_{12}$ e $SARIMA(1,0,0)(1,1,0)_{12}$.

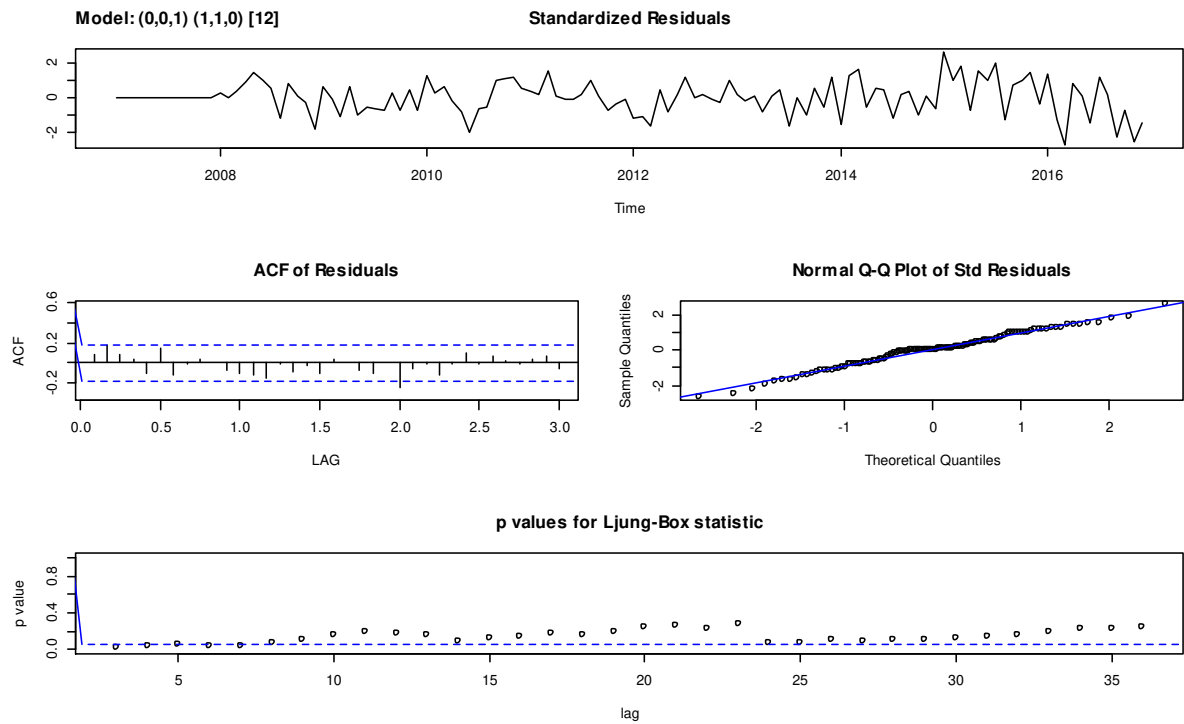


Figura 13. Análise gráfica de resíduos do modelo $SARIMA(0,0,1)(1,1,0)_{12}$.

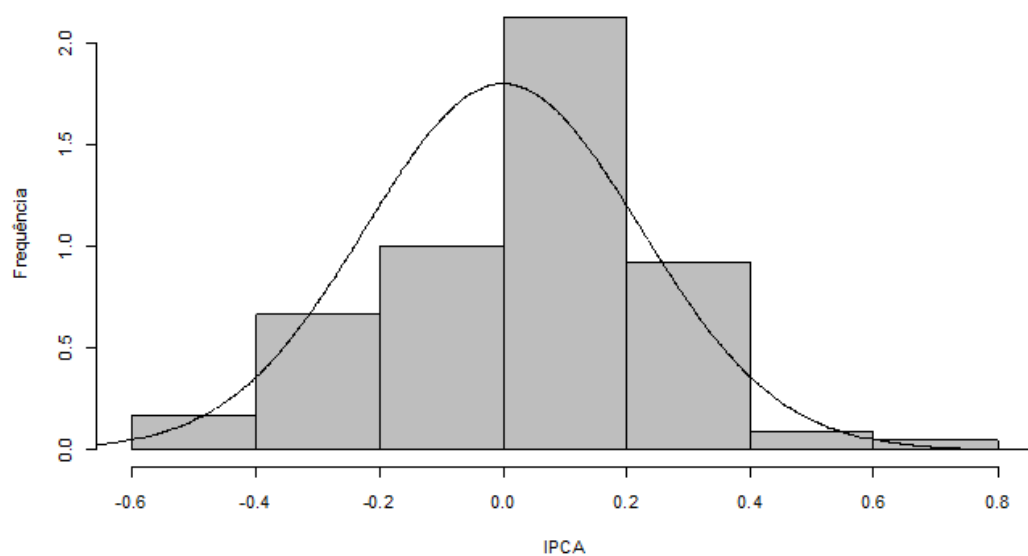


Figura 14. Histograma dos resíduos do modelo $SARIMA(0,0,1)(1,1,0)_{12}$ com curva normal sobrepota.

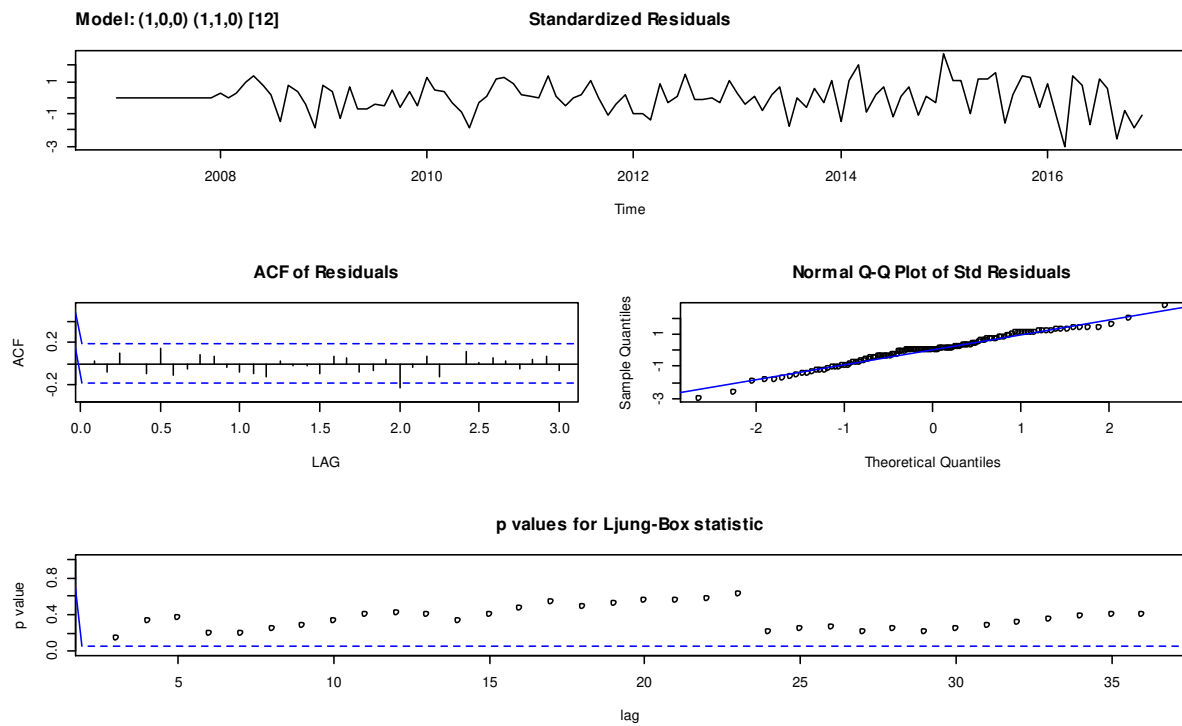


Figura 15. Análise gráfica de resíduos do modelo $SARIMA (1, 0, 0) (1, 1, 0)_{12}$.

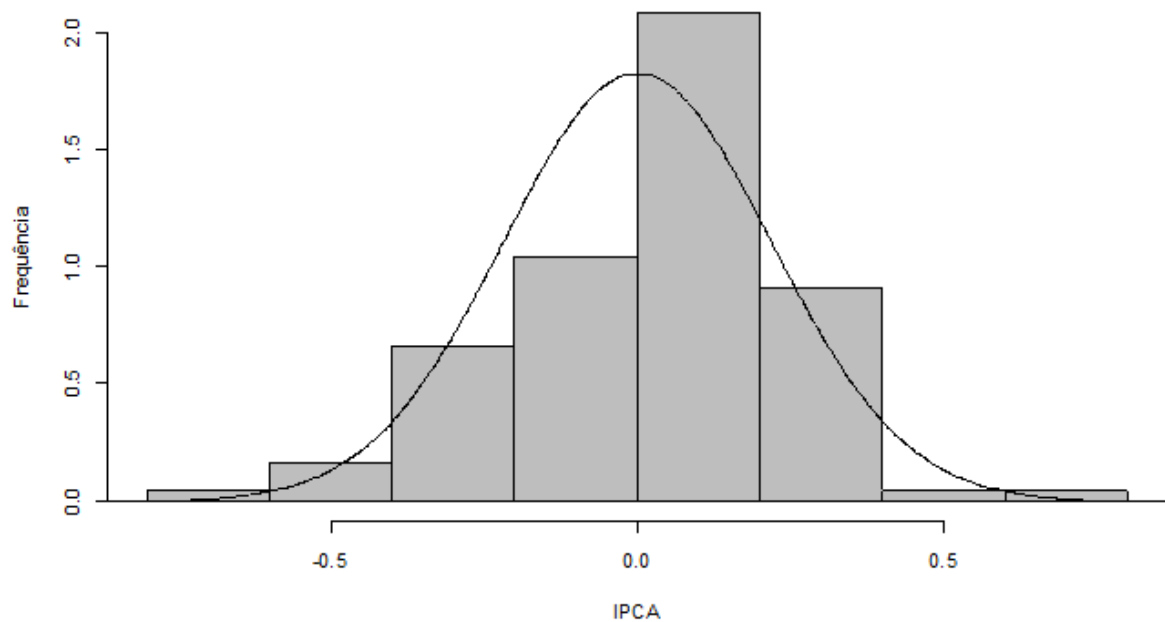


Figura 16. Histograma dos resíduos do modelo $SARIMA (1, 0, 0) (1, 1, 0)_{12}$ com curva normal sobreposta.

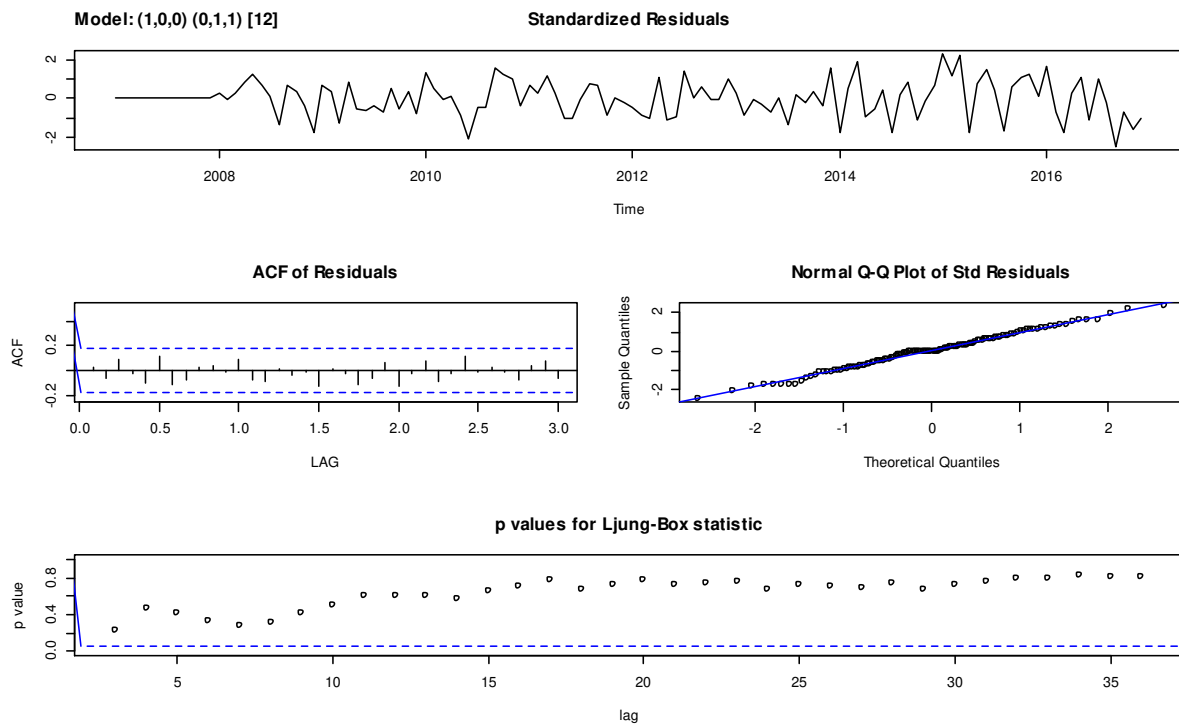


Figura 17. Análise gráfica de resíduos do modelo $SARIMA (1, 0, 0) (0, 1, 1)_{12}$.

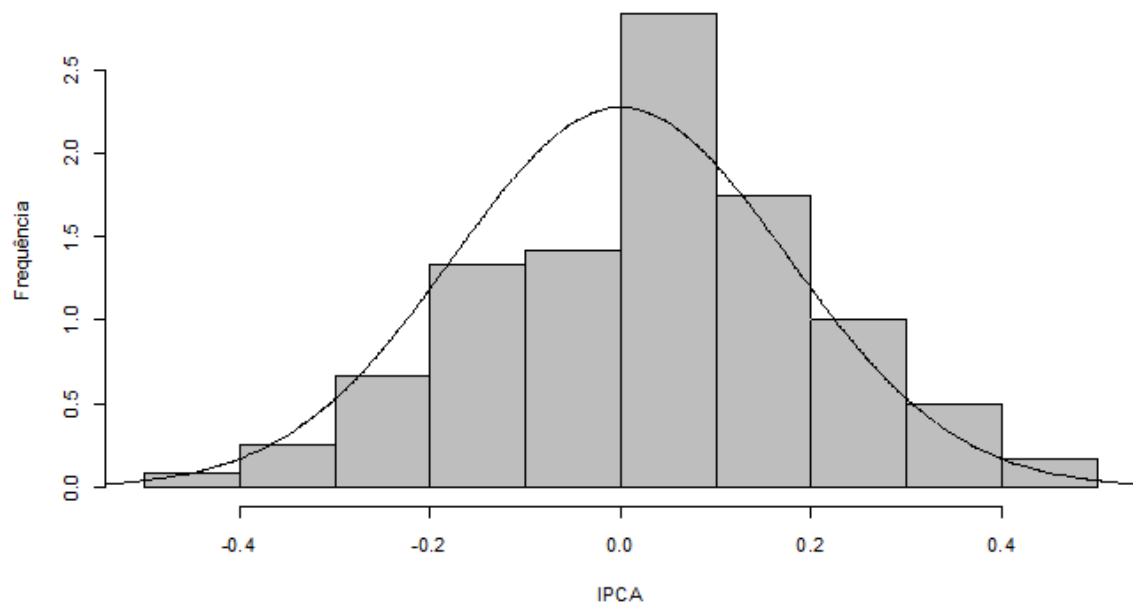


Figura 18. Histograma dos resíduos do modelo $SARIMA (1, 0, 0) (0, 1, 1)_{12}$ com curva normal sobreposta.

APÊNDICE D

Funções e comando R utilizados na análise dos dados.

Pacotes utilizados durante a análise

```
library(agricolae);
library(tseries);
library(TSA);
library(graphics);
library(urca);
library(forecast);
library(GeneCycle);
library(MASS);
library(nortest);
library(astsa);
library(grDevices);
library(PMCMR);
library(ggplot2);
library(seasonal);
library(uroot);
library(randtests);
library(BETS)
```

Definindo a pasta de Trabalho

```
setwd("./"); getwd()
dados<-read.table("")
dados
serie=ts(dados, start=c(2007, 1), end=c(2016, 12), freq=12)
serie
is.ts(serie)
```

Medidas descritivas da série

```
mean(serie)
median(serie)
var(serie)
sd(serie)
range(serie)
kurtosis(serie)
skewness(serie)
sd(serie)/mean(serie)*100
```

Análise gráfica da série original

```
win.graph(width=8,height=5,pointsize=12)
plot(window(serie,start=c(2007,1),end=c(2016,12)),type="l",cex.lab=1.3,lwd=2,
col="red3", xlab="Período (Meses)", ylab="Índice de Preços ao Consumidor Amplo (IPCA)")
Month=c("F"," "," "," "," "," "," "," ","D"); points(window(serie,start=c(2009,1),
end=c(2015, 12)),pch=Month,font=2)
Month=c('J','F','M','A','M','J','J','A','S','O','N','D');
points(window(serie,start=c(2009,1),end=c(2015, 12)),pch=Month,font=2)
abline(v=seq(2009,2017,1),col="grey60",lty=3)
```

Análise da FAC e FACP da série original

```
win.graph(width=10,height=5,pointsize=10); par(mfrow=c(1,2))
a=acf(serie,plot=F,na.action = na.fail,lag.max=36);a$lag; a$lag=a$lag*12;
plot(a,main=NA,col="red",cex.lab=1.3,sub="(1)",cex.sub=1.4,lwd=3,xlab="Lag
(Defasagem)",ylab="Função de Autocorrelação - ACF")
p=pacf(serie,plot=F,na.action = na.fail,lag.max=36);p$lag; p$lag=p$lag*12;
plot(p,main=NA,col="red",cex.lab=1.3,sub="(2)",cex.sub=1.4,lwd=3,xlab="Lag
(Defasagem)",ylab="Função de Autocorrelação Parcial - PACF")
```

Uma transformação nos dados é adequada?

Análise dos dados originais

```
transf<-read.delim("://.txt", header=F); transf
medias=apply(transf,2,mean);
amplitudes=c(apply(transf,2,max))-c(minimo=apply(transf,2,min))
desvios=apply(transf,2,sd); medias; amplitudes; desvios
subconjuntos=cbind(c(medias),c(amplitudes),c(desvios))
colnames(subconjuntos)=c("médias", "amplitudes", "desvios"); print(subconjuntos)
win.graph(width=10,height=5,pointsize=10); par(mfrow=c(1,2))
plot(medias,amplitudes,cex.lab=1.3,sub="(1)",cex.sub=1.4,
pch=19,lwd=1,col="black",xlab="Média",ylab="Amplitude");
modelo<-lm(amplitudes~medias); abline(coef(modelo), col = "red",lwd=1)
plot(medias,desvios,cex.lab=1.3,
sub="(2)",cex.sub=1.4,pch=19,lwd=1,col="black",xlab="Média",ylab="Desvio Padrão");
modelo<-lm(desvios~medias); abline(coef(modelo), col = "red",lwd=1)
```

Análise dos dados com transformação logarítmica

```
x=log(serie)
transflog=read.delim("://.txt", header=FALSE);
transflog; mediaslog=apply(transflog,2, mean)
amplitudeslog=c(apply(transflog,2,max))-c(minimo=apply(transflog,2,min))
desvioslog=apply(transflog,2,sd); mediaslog; amplitudeslog; desvioslog
subconjuntoslog=cbind(c(mediaslog),c(amplitudeslog),c(desvioslog))
colnames(subconjuntoslog)=c("médias", "amplitudes", "desvios"); print(subconjuntoslog)
win.graph(width=10,height=5,pointsize=10); par(mfrow=c(1,2))
plot(mediaslog,amplitudeslog,cex.lab=1.3,sub="(1)",cex.sub=1.4,
pch=19,lwd=1,col="black",xlab="Média", ylab="Amplitude")
modelo<-lm(amplitudeslog~mediaslog); abline(coef(modelo), col = "red",lwd=1)
plot(mediaslog,desvioslog,cex.lab=1.3,sub
="(2)",cex.sub=1.4,pch=19,lwd=1,col="black",xlab="Média", ylab="Desvio Padrão")
modelo<-lm(desvioslog~mediaslog);abline(coef(modelo), col = "red",lwd=1)
```

Análise dos dados com transformação de raiz quadrada

```
x=sqrt(serie)
transfraiz=read.delim("://.txt", header=FALSE);
transfraiz; mediasraiz=apply(transfraiz,2, mean)
amplitudesraiz=c(apply(transfraiz,2,max))-c(minimo=apply(transfraiz,2,min))
desviosraiz=apply(transfraiz,2,sd); mediasraiz; amplitudesraiz; desviosraiz
subconjuntosraiz=cbind(c(mediasraiz),c(amplitudesraiz),c(desviosraiz))
colnames(subconjuntosraiz)=c("médias", "amplitudes", "desvios"); print(subconjuntosraiz)
win.graph(width=10,height=5,pointsize=10); par(mfrow=c(1,2))
```

```
plot(mediasraiz,amplitudesraiz,cex.lab=1.3,sub="(1)",cex.sub=1.4,
pch=19,lwd=1,col="black",xlab="Média", ylab="Amplitude")
modelo<-lm(amplitudesraiz~mediasraiz); abline(coef(modelo), col = "red",lwd=1)
plot(mediasraiz,desviosraiz,cex.lab=1.3,sub="(2)",cex.sub=1.4,pch=19,lwd=1,col="black",
xlab="Média", ylab="Desvio Padrão")
modelo<-lm(desviosraiz~mediasraiz);abline(coef(modelo), col = "red",lwd=1)
```

Análise de sazonalidade e presença de raiz unitária na série

Análise gráfica

Gráfico da série mensal e da média da série mensal

```
win.graph(width=8,height=5,pointsize=10)
monthplot(serie,col="black",ylab="Série IPCA",lwd=1,col.base="red3",lwd.base=2)
legend(locator(1),legend=c("IPCA","Média"),
lty=c(1,1),col=c("black","red3"),cex=0.9,bty="n")
```

Gráfico sazonal anual

```
win.graph(width=8,height=5,pointsize=10)
seasonplot(serie,col="red3")
```

Retorna o gráfico da série decomposta em suas principais componentes

```
win.graph(width=8,height=7,pointsize=10)
plot(decompose(serie,type="additive"),col="red3",lwd=2)
```

Testes formais de sazonalidade

Teste de Fisher para analisar sazonalidade, Periodograma da série

```
win.graph(width=8,height=5,pointsize=10)
pgram=spectrum(serie,fast=FALSE,detrend=FALSE,log="no",plot=T,main="Periodograma
da sérieIPCA",col="red3",lwd=2)$spec
n=length(pgram)
pvalue.g=function(g.obs, n) n*(1-g.obs)^(n-1)
g.obs=max(pgram)/sum(pgram)
pvalue.g(g.obs,n)
pgram2=data.frame(pgram,1:n, p=pgram==max(pgram))
pgram2[pgram2$p==T,]
```

Teste de Kruskal-Wallis para analisar sazonalidade

```
serie=serie
periodo=12
alpha2=0.05
n.serie=length(serie)
# número de grupos
n.grupos = as.integer(n.serie / periodo)
kruskal.g = NULL
sazon = FALSE
# loop gerando os grupos
for (ind.g in 1:n.grupos) {
kruskal.g = c(kruskal.g, rep(ind.g, periodo))
}
# Kruskal-Wallis Rank Sum Test
temppv = kruskal.test (serie, kruskal.g) $ p.value
# p-valor
```

```
sazon.pv = ifelse (!is.na(tempvpv), round(tempvpv,6), 0.5)
```

```
# significância estatística
```

```
sazon = sazon.pv < alpha2 # 90% -> .1
```

```
# tem sazonalidade?
```

```
sazon
```

```
# p-valor
```

```
sazon.pv
```

```
# Testes formais de presença de raiz unitária
```

```
# Retorna o número de diferenças sazonais necessárias na série com base nos testes de raiz unitária de Canova e OSBC. Porém não exclui a necessidade de testes de sazonalidade, uma vez que uma série pode apresentar padrão sazonal sem, no entanto, possuir raiz unitária sazonal.
```

```
nsdiffs(serie, m = frequency(diff(serie)), test = c("ocsb", "ch"), max.D = 1)
```

```
# Teste de Canova para período específico que aparenta apresentar sazonalidade
```

```
ch.test(serie, type = "dummy", sid=7)
```

```
# Teste de Canova por mês e geral
```

```
ch.test(serie)
```

```
# Outras ferramentas auxiliares
```

```
# Propõe um modelo ARIMA automático para a série. Pode ser utilizado para fins de comparação.
```

```
auto.arima(serie)
```

```
# Faz o ajuste sazonal da série. Retorna a transformação utilizada nos dados, e o modelo sazonal ajustado. Pode ser utilizado para fins de comparação.
```

```
ajuste=seas(serie);summary(ajuste)
```

```
# Retorna o melhor modelo de alisamento exponencial para os dados, permite verificar a natureza das componentes da série (se a tendência, erro e sazonalidade estão presentes na série e se são aditivos ou multiplicativos).
```

```
saz=ets(serie); summary(saz)
```

```
##### Diferenciação Sazonal da Série #####
```

```
# 1ª diferença sazonal da série
```

```
difs=diff(serie, lag=12)
```

```
difs
```

```
##### Análise da Tendência da Série #####
```

```
# Testes formais para Tendência e raiz unitária
```

```
# Teste de Cox-Stuart para tendência
```

```
cox.stuart.test(difs)
```

```
# Testes de presença de raiz unitária
```

```
# Retorna o número de diferenças simples necessárias na série com base nos testes de raiz unitária ADF, PP e KPSS.
```

```
ndiffs(difs, alpha = 0.05, test = c("kpss", "adf", "pp"), max.d = 2)
```

Teste ADF

```
adf.test(difs)
```

Teste PP

```
pp.test(difs)
```

Teste KPSS

```
kpss.test(difs)
```

Identificação e Estimação dos modelos

```
x=serie
```

Modelo 1

```
m1=arima(x, order = c(0,0,1),seas = list(order = c(1,1,0), period=12),method="ML")
win.graph(width=8,height=5,pointsize=10); tsdiag(m1); t(confint(m1)); res=m1$residuals
shapiro.test(res); summary(m1)
```

Modelo 2

```
m2=arima(x, order = c(1,0,0),seas = list(order = c(1,1,0), period=12),method="ML")
win.graph(width=8,height=5,pointsize=10); tsdiag(m2); t(confint(m2)); res=m2$residuals
shapiro.test(res); summary(m2)
```

Modelo 3

```
m3=arima(x, order = c(1,0,0),seas = list(order = c(0,1,1), period=12),method="ML")
win.graph(width=8,height=5,pointsize=10); t(confint(m3)); tsdiag(m3); res=m3$residuals
shapiro.test(res); summary(m3)
```

Análise gráfica de resíduos e critérios AIC AICC BIC**# Modelo 1**

```
win.graph(width=8,height=5,pointsize=10)
sarima(x,0,0,1,1,1,0,12)
```

Modelo 2

```
win.graph(width=8,height=5,pointsize=10)
sarima(x,1,0,0,0,1,1,12)
```

Modelo 3

```
win.graph(width=8,height=5,pointsize=10)
sarima(x,1,0,0,1,1,0,12) # Modelo m3
```

Histograma dos resíduos com curva normal sobreposta**# Modelo 1**

```
win.graph(width=8,height=5,pointsize=10)
z = m1$residuals
d = seq(range(z)[1] - 3 * sd(z), range(z)[2] + 3 * sd(z), 0.001)
hist(z, freq = F,xlab="IPCA",ylab="Frequência",col="grey",main="")
lines(d, dnorm(d, 0, sd(z)))
```

Modelo 2

```
win.graph(width=8,height=5,pointsize=10)
z = m2$residuals
d = seq(range(z)[1] - 3 * sd(z), range(z)[2] + 3 * sd(z), 0.001)
```



```
hist(z, freq = F,xlab="IPCA",ylab="Frequência",col="grey",main="")
lines(d, dnorm(d, 0, sd(z)))
```

Modelo 3

```
win.graph(width=8,height=5,pointsize=10)
z = m3$residuals
d = seq(range(z)[1] - 3 * sd(z), range(z)[2] + 3 * sd(z), 0.001)
hist(z, freq = F,xlab="IPCA",ylab="Frequência",col="grey",main="")
lines(d, dnorm(d, 0, sd(z)))
```

Testes formais sobre os resíduos dos modelos estimados

```
m= m1
```

Testes de independência dos resíduos

Ljung Box

```
Box.test(m$residuals, lag=24, type = "Ljung-Box", fitdf=2)
```

Box-Pierce

```
Box.test(m$residuals, lag=24, type = "Box-Pierce", fitdf=2)
```

Testes de normalidade dos resíduos

Shapiro-Wilk

```
shapiro.test(m$residuals)
```

Kolmogorov-Smirnov

```
ks.test(m$residuals, "pnorm", mean(m$residuals), sd(m$residuals))
```

Jarque-Bera

```
jarque.bera.test(m$residuals)
```

Previsões $h = 8$

```
p = as.data.frame(predict(m,8, interval="prediction")); p
```

Intervalo de confiança 95%

```
LI=p$pred - 1.96*p$se; LS=p$pred + 1.96*p$se
```

Período, Limite Inferior, Previsões ($h = 8$), Limite Superior

```
p=p$pred
```

```
Per=c((length(serie)+1):(length(serie)+8))
```

```
df=data.frame(Per,LI,p,LS)
```

```
print(df,row.names=FALSE)
```

Comparação gráfica entre os três modelos aptos

```
teste=read.table("://.txt")
```

```
seriereal=ts(teste,start=c(2017,1),end=c(2017,8),freq=12)
```

```
seriereal
```

```
a = as.data.frame(predict(m1,8, interval="prediction")); p1=a$pred
```

```
b = as.data.frame(predict(m2,8, interval="prediction")); p2=b$pred
```

```
c = as.data.frame(predict(m3,8, interval="prediction")); p3=c$pred
```

```
mes=c(1:8)
```

```
obs=c(seriereal)
```

```
e1=c(p1)
```

```
e2=c(p2)
```

```
e3=c(p3)
```

```

win.graph(width=8,height=5,pointsize=12)
plot(mes,obs,cex.lab=1,pch=16,type="o",col="#67000D",lwd=3,
xlab="Período (janeiro de 2017 a agosto de 2017)",
ylab="Índice de Preços ao Consumidor Amplo (IPCA)",ylim=c(-1,2))
lines(mes,e1,col="#FB6A4A",lwd=3,type="o",pch=16)
lines(mes,e2,col="#CB181D",lwd=3,type="o",pch=16)
lines(mes,e3,col="#FCBBA1",lwd=3,type="o",pch=16)

legend(locator(1), xpd=TRUE, ncol=1, legend=c("Série de valores reais","SARIMA (0,0,1)
(1,1,0)","SARIMA (1,0,0) (0,1,1)","SARIMA (1,0,0) (1,1,0)"),
fill=c("#67000D","#FB6A4A","#CB181D","#FCBBA1"),bty="n")

```

Gráfico do ajuste do modelo m2 (SARIMA (1, 0, 0) (0, 1, 1)₁₂)

```

win.graph(width=8,height=5,pointsize=12)
m2=arima(x, order = c(1,0,0),seas = list(order = c(0,1,1), period=12),method="ML")
fitm2=fitted.Arima(m2)
plot(fitm2,xlab="Período (Meses)", ylab="Índice de Preços ao Consumidor Amplo (IPCA)",
col="red3",ylim=c(-0.5,1.5))
lines(serie,col="black")
legend(locator(1),legend=c("Série IPCA","Modelo ARIMA (1, 0, 0) (0, 1, 1)"),
lty=c(1,1),lwd=c(2,2),col=c("black","red3"),cex=0.8,bty="n")

```

Gráfico do IPCA + Valores Previstos + Intervalos de Confiança

```

win.graph(width=8,height=5,pointsize=12)
plot(window(serie, start=2007),
xlim=c(2007, 2018.1),ylim=c(-0.5,1.5),
col="red3", lwd=2, ylab="Índice de Preços ao Consumidor (IPCA)")
lines(p$pred,col="darkred",lwd=2)
abline(v=seq(2017,2017,1),col="grey1",lty=3)
lines(p$pred+1.96*p$se,col="darkred",lwd=1,lty=2)
lines(p$pred-1.96*p$se,col="darkred",lwd=1,lty=2)
lines(seriereal,col="black",lwd=1)
legend(locator(1),legend=c("Série IPCA","ValoresPrevistos","Intervalo de Confiança"),
lty=c(1,1,2),lwd=c(2,2,1),col=c("red","darkred","darkred"),cex=0.8,bty="n")

```